HieraSurg: 分层感知扩散模型用于手术视频生成

Diego Biagini[1,2], Nassir Navab[1,2], and Azade Farshad[1,2]

[1]Chair for Computer Aided Medical Procedures (CAMP), TU Munich, Germany
[2]Munich Center for Machine Learning (MCML)
{diego.biagini,azade.farshad}@tum.de

摘要 手术视频合成在扩散模型在通用领域视频生成的成功之后，成为了一个有前景的研究方向。尽管现有方法实现了高质量的视频生成，但大多数都是无条件的，并且无法保持与手术动作和阶段的一致性，缺乏实现事实模拟所需的手术理解和精细指导。我们通过提出 **HieraSurg** 来解决这些挑战，这是一个层次感知的手术视频生成框架，由两个专门的扩散模型组成。给定一个手术阶段和初始帧，HieraSurg 首先通过分割预测模型预测未来的粗粒度语义变化。然后通过第二阶段模型增强这些时态分割图以添加细粒度视觉特征来生成最终视频，从而实现有效的纹理渲染并在视频空间中整合语义信息。我们的方法利用了多个抽象层次的手术信息，包括手术阶段、动作三元组和全景分割图。在胆囊切除术视频生成上的实验结果表明，模型在定量和定性上均显著优于先前的工作，显示出强大的泛化能力和生成更高帧率视频的能力。当提供现有的分割图时，该模型表现出特别精细的一致性，这表明其具有实际手术应用的潜力。
项目网页：diegobiagini.github.io/HieraSurg/

1 介绍

模型效率、扩散建模和数据整理方面的最新进展为生成高度逼真的视频铺平了道路 [1], [2], [3]。这些发展表明，此类模型不仅仅是复制数据驱动的相关性；相反，它们建立了一个内部世界模型，展现出类似于模拟器的新兴能力 [4]。

在医学领域，一个自然产生的问题是：视频生成模型是否能够在手术这种由隐含规则支配场景演变的受限环境中可靠地运行？在这项工作中，我们评估了这些模型作为预测工具的潜力，以预期手术场景的短期演变。此外，本研究强调了生成模型在手术数据科学中的另一个重要应用，即缓解低数据环境下的数据稀缺问题 [5]。先前探索手术领域中生成视频模型的工

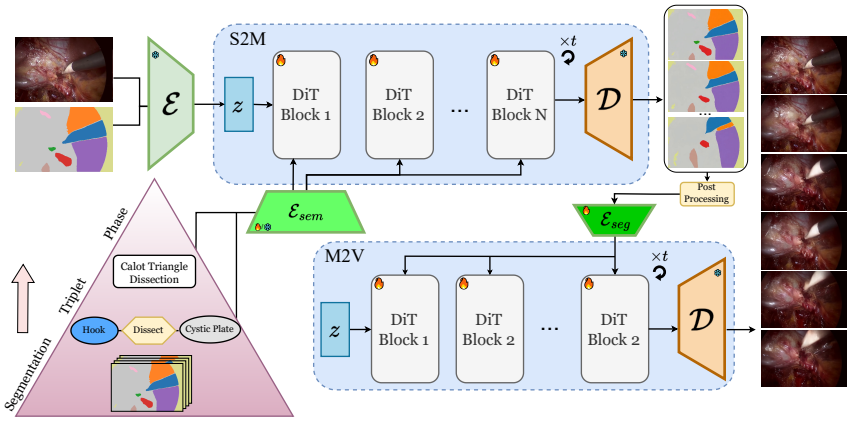

图 1. HieraSurg 管道流程. **左**: 我们采用手术场景的层级表示，**正确**: HieraSurg 的推理管道和组件。S2M 接受初始分割图，并通过考虑阶段和三元组信息来预测手术场景的未来发展。最后，S2M 的输出被送入 M2V 以根据预测的分割生成视频。

作包括 Endora[6]，这是一个无条件的视频模型，在训练过程中整合了来自 DINO[7] 主干网络的语义特征，以及 VISAGE[8]，其中视频生成是基于通过 CLIP[9] 编码器嵌入的动作三元组进行条件化。然而，大多数这些工作只考虑了手术场景的一个方面。特别是，我们认为可以以分层的方式描述一个手术场景 (参见 Figure 1)，其中信息可以在不同的抽象层次上被结构化。我们提出，生成过程应当沿金字塔的垂直轴分割，某些组件应针对层级中的特定层进行调整。这是背后的核心思想**希拉手术**及其组成部分：一个第一阶段模型 **HieraSurg-S2M（语义到地图）**，接收场景的高层次手术信息以生成一组全景分割图，这些图像用于指导第二阶段模型**希拉手术-M2V（映射到视频）**生成视频。在训练第二阶段模型时，需要大量的分割图来学习语义空间和视频空间之间的关系。由于数据稀缺，我们基于 Segment Anything 2 (SAM2)[10] 设计了一个自动化标注流水线，从未标记的手术视频中提取全景分割图。

我们验证了在来自 Cholec80[11] 和 CholecT45[12] 数据集的腹腔镜手术中的框架。我们的主要贡献如下：1)**HieraSurg**，一对耦合的生成式视频模型，可以合成具有前所未有的视觉质量和帧率的真实腹腔镜视频；2) 一个专为低帧率外科视频设计的自动全景分割流水线；3) 使用保真度和重建/检测为基础的指标对 HieraSurg 的视觉质量、运动忠实度以及短期场景预测性能进行了广泛的实验评估。

2 方法

定义在一个给定的视频数据集中，每个视频是一个由 F 帧组成的序列 $V = x_i, ..., x_F$，每帧的分辨率为 H 乘以 W。我们将编码器映射一个视频到低维潜在空间表示为 $\mathcal{E}$，这是一种遵循潜扩散模型（LDMs）的变分自编码器（VAE）[13]。从潜在空间回到图像空间的映射由解码器 $\mathcal{D}$ 执行。视频的全景分割图是一个张量 $x \in \mathbb{N}^{F \times H \times W}$，其中每个像素被分配一个整数值或背景值，表示该处发现的对象实体。从此以后，我们将把全景分割图简称为分割图。

扩散模型 扩散模型通过逐步去噪高斯样本生成数据，以匹配给定分布，这一过程通过前向和后向扩散实现。前向过程在数据样本 $\boldsymbol{x}_0$ 上添加高斯噪声，经过 T 个时间步长，产生一系列逐渐变噪的样本序列 $\boldsymbol{x}_1, ..., \boldsymbol{x}_T$。在每个时间步 t，根据 $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I})$ 添加噪声，其中 $\beta_{t\,t=1}^{T}$ 是一个固定的噪声调度，$\mathcal{N}(\mu, \sigma^2)$ 表示均值为 μ 和方差为 σ^2 的高斯分布。逆过程旨在通过计算后验 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 来学习如何去噪，这在实际上是不可行的，并且需要近似方法。在 [14] 中引入的标准公式中，神经网络 ϵ_θ 预测每一步的噪声成分。逆过程可以表示为：

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I})$$

其中 $\mu_\theta(\boldsymbol{x}_t, t)$ 由 $\epsilon_\theta(x_t, t)$ 派生。去噪模型通过优化对数似然的一个变分下界进行训练，简化了目标：

$$\mathcal{L} = \mathbb{E}_{t, \boldsymbol{x}_0, \epsilon}\left[|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t)|_2^2\right]$$

其中，ϵ 是正向过程中添加的随机噪声，而 $\boldsymbol{x}_t$ 由以下方式获得：$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon$ 带有 $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$。在采样过程中，通过从随机高斯噪声 $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ 开始，并迭代应用学习到的逆过程 T 步，预测并去除噪声成分，直到获得干净样本 $\boldsymbol{x}_0$。

视频分割 在我们的视频生成流程中，我们严重依赖分割图作为中间表示。获得此类手术数据的标签成本很高，原因是记录长度较长，现有的数据集如 CholecSeg8k[15] 和 Endoscapes[16] 都过于有限。鉴于扩散模型的数据密集特性，需要一个自动标注流程。尽管一些工作已经解决了这个问题 [17]，但没有一项完全满足我们的特定需求。因此，我们开发了一个利用通用分割模型——即 SAM 和 SAM2[10], [18]——以及稳健特征提取器的管道。最初，我们使用 DINO[7] 作为我们的特征提取器；然而，由于 RADIO[19] 对

非正方形图像和与 SAM 更兼容的特征表示的原生支持，它产生了更好的结果。对于每个帧，我们在网格状模式下采样点来提取分割图，这些样本充当 SAM2 的提示。对于每一个被分割的对象，我们计算 RADIO 特征，并对所有帧重复这个过程。当检测到新的分割时，将其 RADIO 特征与前一帧的特征进行比较；如果距离低于阈值，则将这些实体标记为在不同帧之间移动的同一对象，并记录帧号和初始位置。然后使用此信息运行每个对象的 SAM2 视频跟踪管道，其初始位置作为提示。最后，通过合并重叠的实体并用最具代表性的实体替换经常被覆盖的实体来对分割图进行后处理。尽管为每个实体执行分割以及后续后处理带来了一定的开销，但我们的方法显著优于传统的 SAM2 方法——即同时提示所有对象的方法。

2.1 HieraSurg

我们将 HieraSurg 定义为基于 CogVideoX-2B[2] 架构的一对扩散模型——潜在扩散变压器 (DiT)[20]——该模型接受文本提示以生成不同分辨率和帧率的视频。HieraSurg 由以下部分组成：HieraSurg-S2M（语义到映射），负责生成手术场景在分割图领域中的合理演化，以及 HieraSurg-M2V（映射到视频），完成任务即将所述的分割图带入视频空间。

HieraSurg-S2M：此模型生成一系列 F 分割图。它将第一个视频帧 y_1、其对应的分割图 y_1^{seg} 以及接下来的 F 个阶段 ph_i 和动作三元组 tr_i 作为输入。我们不通过标准编码器 $\mathcal{E}$ 来对视频进行编码，而是执行时间密集型潜在编码以保留单帧细节。我们通过将一个视频视为一批 F 单帧视频并分别对它们进行编码来实现这一点，从而获得潜在的 $z \in \mathbb{R}^{F\times H'\times W'\times d}$，其中帧维度是 F 而不是 F'，F' 是使用 $\mathcal{E}$ 编码长度为 F 的视频时潜变量的时间维度。虽然每个分割图自然是单通道的二维表示，但在将它们输入 VAE 之前，我们将它们转换为颜色空间，以便其分布更好地匹配 CogVideoX 的预训练数据。我们在两个轴上注入条件，通过 VAE 对初始分割帧和第一个视频帧进行编码，分别表示为 $z_y^{seg} = \mathcal{E}(y_1^{seg}), z_y = \mathcal{E}(y_1) \in \mathbb{R}^{1\times H'\times W'\times d}$，并在帧维度上重复这些以匹配 z。最终将它们拼接并作为去噪模型 ϵ_θ 的输入提供如下：

$$z_{in} = [z, z_y^{seg}, z_y] \in \mathbb{R}^{F\times H'\times W'\times 3d}$$

为了对 ph_i 和 tr_i 进行条件设置，我们分别对每个编码，然后使用一维卷积后跟平均池化沿时间维度压缩它们。我们尝试了一个可学习嵌入层（标签嵌入）以及一个预训练模型如 PeskaVLP[21]，在这些模型中，相位和三元组首

先被映射到文本表示形式然后再进行嵌入。生成的编码相位和三元组信息然后与时间步长 t 的编码拼接，并注入到每个变压器块中。扩散模型在连续空间中运行，这意味着 S2M 会产生具有略微不同值的分割图——这对于精确分割来说是一个不希望的结果。为了解决这个问题，我们对输出颜色集应用 K-Means 聚类算法，使用肘部法算法找到最优的聚类数量。然后我们将每个聚类内的所有颜色映射到它们的中心，从而得到一个离散表示，该表示可以很容易地映射到单通道整数空间。

HieraSurg-M2V： 给定输入分割图 c 和视频的第一帧 y_1，生成一个时长变化的视频序列。我们允许分割图具有与输出视频相同数量的帧，或者在生成高帧率视频的情况下拥有较少的数量。为了编码分割图，我们通过一个由一系列 3D 残差块组成的编码器 $\mathcal{E}_{seg}$ 进行处理，每个块都会降低空间分辨率而保持时间维度不变。

我们进一步向仍为空间压缩的 3D 空间 $\mathcal{E}_{seg}(c)$ 添加了具有时间感知的正弦位置嵌入，在将表示展平为长度为 T_{seg} 的一维序列 H_0^{seg} 之前。H_0^{seg} 通过每个变压器块中的注意力机制合并到网络的主流程中，在那里进一步处理：

$$H_{cat} = [H_i; H_i^{seg}] \qquad Q = W_q H_{cat} \quad K = W_k H_{cat} \quad V = W_v H_{cat}$$

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \qquad [H_{i+1}, H_{i+1}^{seg}] = \text{split}(W_o Z, T)$$

其中 H_{cat} 是经过块 i 的隐藏状态 H_i 和 H_i^{seg} 的连接，d_k 是键向量的维度，而 W_q, W_k, W_v, W_o 是可学习的权重矩阵。类似于 S2M，我们将初始视频帧作为单个图像通过 VAE 编码，然后再与其噪声潜变量堆叠在一起。

3 实验

数据集和预处理：我们使用了 Cholec80 数据集 [11]，这是一个由 13 名外科医生进行的 80 次胆囊切除手术组成的集合，以每秒 25 帧的高分辨率视频形式获取，并配有阶段标注；以及 CholecT45 数据集 [12]，这是其中的一个子集，包含 45 条以每秒 1 帧的速度进行注释的三元类。提取了 65 段视频（其中有 41 段来自 CholecT45），并对以每秒 1 帧采样的重叠 16 秒长片段运行自动分割管道。两个视频被保留作为测试集。我们将原始视频从 854x480 裁剪并调整大小到 384x256。

评估设置 我们使用 Frechet 视频距离（FVD）[22]、Frechet Inception Distance (FID)[23] 和使用 PeskaVLP[21] 作为特征提取器的 FID 来评估生成的视频

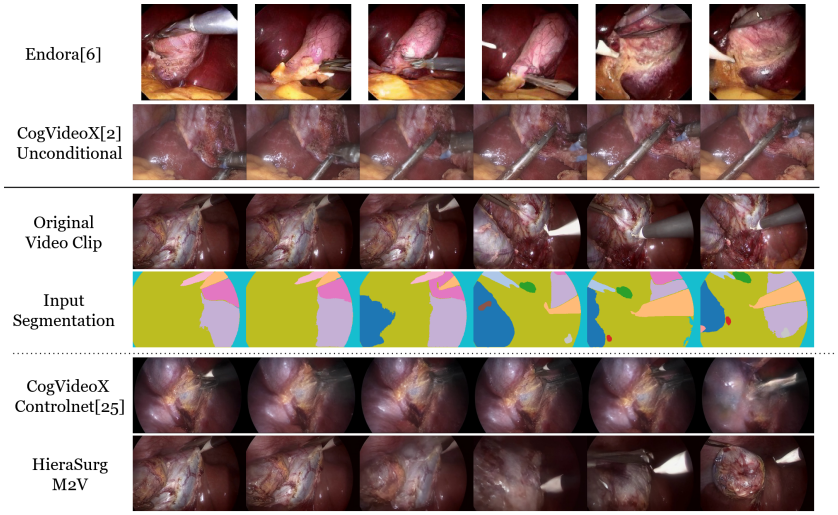

图 2. 不同模型生成的前 6 秒的视觉比较。对于条件模型，输入分割图已给出。

的视觉质量。重建能力使用结构相似性指数测量（SSIM）进行评估。度量是在 1024 个生成样本与相同数量的真实数据点之间计算的。为了验证分割一致性，我们引入了一个检测器一致性指标。训练一个 YOLOV8[24] 模型来识别手术工具后，在两个视频的所有帧上运行工具检测。通过匹配边界框预测，我们计算在真实视频中找到的生成视频中的对象比例（命中率真实），在真实视频中找到的生成对象的比例（命中率生成）以及匹配预测的平均 IoU。

结果我们首先验证 HieraSurg-M2V 合成 1FPS 16 秒视频的能力。在通过训练清零后的字幕来微调 CogVideoX-2B 文本到视频模型以生成无条件的手术视频后，得到的权重被用作所有 M2V 变体的起点。我们实验了通过不同的编码方法（要么是 RADIO，要么基于 VAE）为 M2V 提供初始图像。我们在 8FPS 6 秒的视频上进一步训练和评估表现最佳的模型。M2V 的条件跟随能力与流行的注入控制到 LDM 架构中的 ControlNet[25] 进行了比较。特别是，我们采用了为 CogVideoX 实施的使用 Canny 边缘的 ControlNet，并在从我们的分割映射中提取的边缘上对其进行微调。结果显示（参见 Table 1），通过 M2V 在保真度指标上的改进值尤为显著，在 8FPS 设置下尤为如此。与基于 VAE 编码提供初始帧的基线 ControlNet 相比，我们方法遵循指导的能力明显更优。可视化结果见 Figure 2。接下来，我们在 1FPS 下训练 HieraSurg-S2M，允许预测未来 16 秒的分割图。我们尝试使用预训练和学习

表 1. 希拉手术与无条件和有条件基线的定量比较。S2M+M2V 表示完整的 HieraSurg 管道。* 报告在论文中的值。+Cholec80 模型在我们的测试划分上进行评估。

Model	Conditioning	保真度度量标准 (↓)			重建度量 (↑)			
		FVD	FID	PeskaVLP FID	HR Real	HR Gen	MIoU	SSIM
帧率: 1								
Endora [6]+	-	815.8	105.2	60.3	-	-	-	0.16
CogVideoX [2]	-	443.1	79.6	35.5	0.007	0.016	0.596	0.28
VISAGE [8]*	Triplet	1780	-	-	-	-	-	0.56
CogVideoX ControlNet [25]	GT Seg. Edges	640.9	82.6	28.7	0.028	0.061	0.616	0.38
M2V RADIO	GT Seg.	640.8	72.7	36.3	0.207	0.406	0.682	0.40
M2V VAE	GT Seg.	351.4	48.9	22.1	**0.321**	**0.476**	**0.745**	**0.49**
S2M+M2V VAE	Pred Seg.	**312.4**	**47.1**	**17.2**	0.137	0.270	0.697	0.44
帧率: 8								
M2V RADIO	GT Seg.	1202.4	71.7	33.5	0.076	0.157	0.635	0.35
M2V VAE	GT Seg.	**276.2**	**22.4**	**12.9**	**0.358**	**0.447**	0.806	**0.53**
S2M+M2V VAE	Pred Seg.	278.0	24.1	15.1	0.218	0.311	**0.867**	**0.53**

到的嵌入来编码阶段和三元组信息。在 Table 2中，我们展示了前者方法给出更好的结果。因此，在后续实验中使用了这种方法。通过在 1FPS 和 8FPS 设置下将 S2M 的输出提供给 M2V，验证生成预测的一致性。当使用预测的分割图时，我们注意到视觉质量仅受到轻微影响，而跟随手术真实世界演变的能力有所下降，这是预期之中的。在 1FPS 设置下检测器协议显著降低，这是因为与预测 6 秒相比，预测未来 16 秒固有的复杂性更大。完整的定量评估可以在 Table 1中找到，而整个流水线的视觉比较则在 Figure 3中给出。S2M 的消融研究在 Table 2中提供，表明手术信息的存在确实有所帮助，并且如 PeskaVLP 所提供的那样，提供稳健的文本编码对阶段和三元组是有益的，优于让模型从头开始学习它们。此外，我们通过展示时间上压缩的潜在空间会生成更难解析的分割图来验证隐式编码的选择。最值得注意的是，消

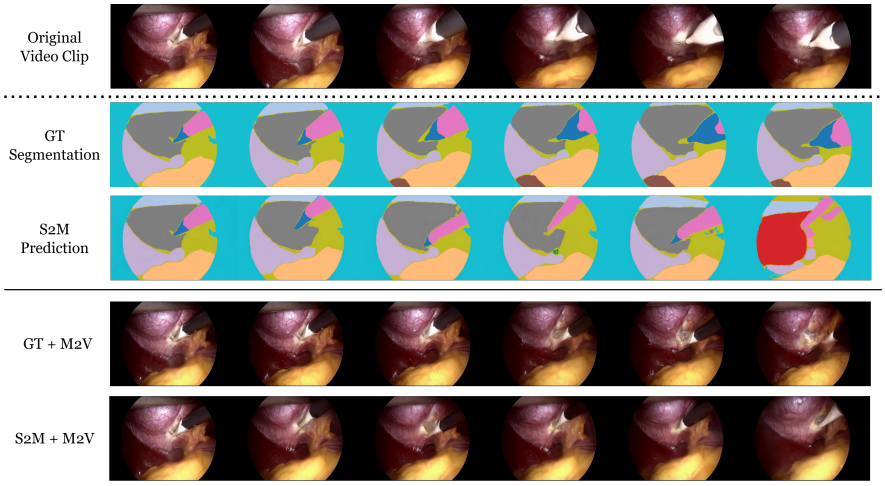

图 3. 示例输出完整的 HieraSurg 管道，比较使用 M2V 与地面真实分割的情况，以及当提供 S2M 输出时的情况。

表 2. S2M 的消融研究，使用完整的 HieraSurg 管道@1FPS

Temporally Dense Latent	Phase/Triplet Cond.	Cond Method	FVD (↓)	FID(↓)	HR Real (↑)	SSIM (↑)
✓	✗	-	378.6	52.5	0.113	**0.44**
✗	✓	PeskaVLP	470.3	50.4	0.104	0.38
✓	✓	Label Emb.	389.6	53.2	0.119	0.43
✓	✓	PeskaVLP	**312.4**	**47.1**	**0.137**	**0.44**

融的第一阶段输出显示边缘模糊，并且在本应均匀的区域中颜色变化更大，这使得 K-Means 过程的效果较差。

4 结论

本工作提出了 HieraSurg 用于高质量和可控的手术视频生成。HieraSurg-M2V 能够合成具有令人印象深刻的视觉质量的视频，严格遵循通过分割图提供的条件。得益于 HieraSurg-S2M，我们进一步能够仅从初始帧和手术阶段/三元组信息生成完全新颖的视频，为场景可能的发展提供了一瞥。这为在多种方法中的应用铺平了道路，既可以作为 M2V 来可视化已知场景的发展，也可以作为完整的 HieraSurg 用于自由形式生成任务。然而，我们的管道预测能力严重依赖于分割图的质量以及第一阶段组件跟踪并贡献有效轨

迹的能力。进一步的开发将集中在改进第一阶段模型上，理想情况下利用更多的语义信息，这可以限制可想象的轨迹范围，使对未来场景状态的更可靠预期成为可能。

References

[1] A. Blattmann et al., "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[2] Z. Yang et al., "Cogvideox: Text-to-video diffusion models with an expert transformer," in *The Thirteenth International Conference on Learning Representations*, 2025.

[3] W. Kong et al., *HunyuanVideo: A systematic framework for large video generative models*, Jan. 17, 2025.

[4] B. Kang et al., "How far is video generation from world model: A physical law perspective," *arXiv preprint arXiv:2411.02385*, 2024.

[5] D. G. Saragih, A. Hibi, and P. N. Tyrrell, "Using diffusion models to generate synthetic labeled data for medical image segmentation," *International journal of computer assisted radiology and surgery*, vol. 19, no. 8, pp. 1615–1625, 2024.

[6] C. Li et al., "Endora: Video generation models asăendoscopy simulators," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru et al., Eds., Cham: Springer Nature Switzerland, 2024, pp. 230–240.

[7] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[8] Y. Yeganeh et al., "VISAGE: Video synthesis using action graphs forăsurgery," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024 Workshops*, M. E. Celebi, M. Reyes, Z. Chen, and X. Li, Eds., Cham: Springer Nature Switzerland, 2025, pp. 146–156.

[9] A. Radford et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

[10] N. Ravi et al., *SAM 2: Segment anything in images and videos*, Oct. 28, 2024.

[11] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on

laparoscopic videos," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, Jan. 2017, Conference Name: IEEE Transactions on Medical Imaging.

[12] C. I. Nwoye et al., "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Medical Image Analysis*, vol. 78, p. 102 433, May 1, 2022.

[13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10 684–10 695.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[15] W.-Y. Hong, C.-L. Kao, Y.-H. Kuo, J.-R. Wang, W.-L. Chang, and C.-S. Shih, *CholecSeg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80*, Dec. 23, 2020.

[16] A. Murali et al., "The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark," *arXiv preprint arXiv:2312.12429*, 2023.

[17] Y. Li, H. Ling, I. V. Ramakrishnan, P. Prasanna, A. Sasson, and H. Gupta, "Critical view of safety assessment in laparoscopic cholecystectomy via segment anything model," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–6.

[18] A. Kirillov et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[19] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov, "AM-RADIO: Agglomerative vision foundation model reduce all domains into one," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12 490–12 500.

[20] W. Peebles and S. Xie, "Scalable diffusion models with transformers," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4195–4205.

[21] K. Yuan, V. Srivastav, N. Navab, and N. Padoy, "Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation," presented at the The Thirty-eighth Annual Conference on Neural Information Processing Systems, Nov. 6, 2024.

[22] T. Unterthiner, S. v. Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "FVD: A new metric for video generation," Apr. 19, 2019.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[24] R. Varghese and S. M., "YOLOv8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Apr. 2024, pp. 1–6.

[25] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 3836–3847.