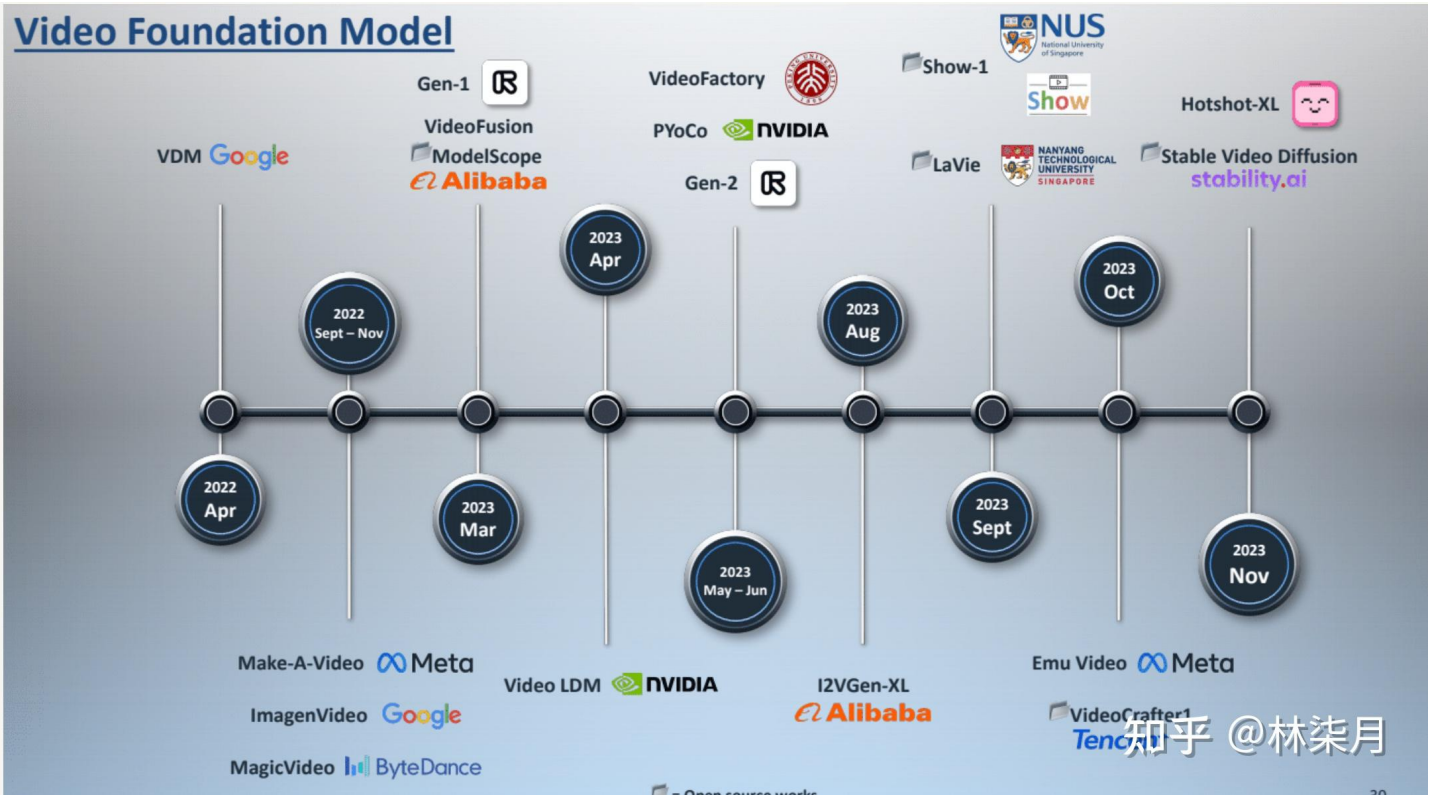
Video Diffusion Models

终于对视频领域的生成模型有个全面的了解了，虽然很多具体的概念还不太明白，但算是心里有一块大地图了吧。

最近学习了NUS Showlab出品的视频生成tutorial，这里全面的带我们过了一遍最近一年也就是从22年初开始视频领域扩散模型的主要进展和大小方向，感觉受益良多。为了让我更好的记住这些精华内容，于是就有了这篇学习笔记 。



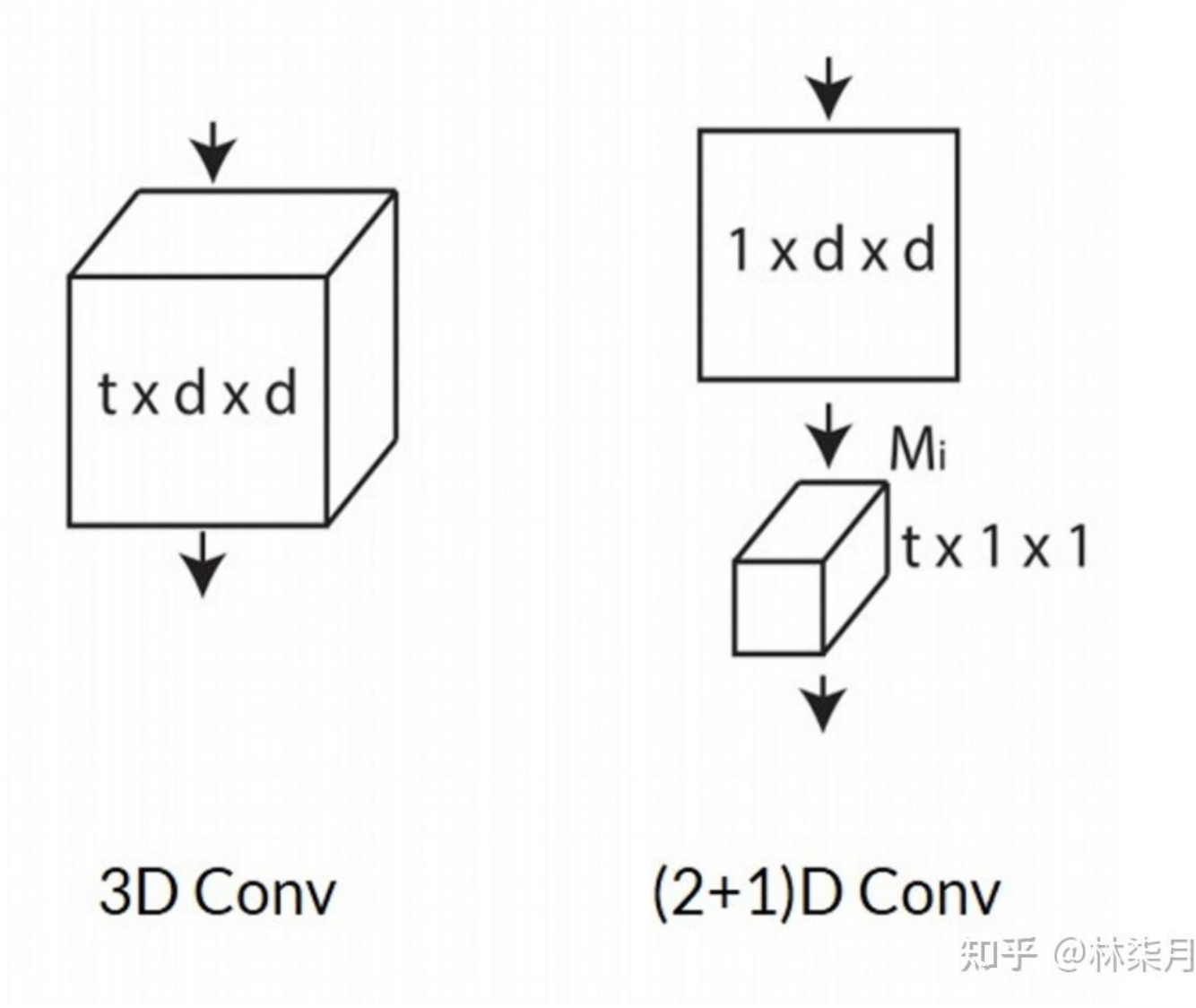
整体发展脉络

在扩散模型出来以后，几乎所有的图像生成大模型都是在采用扩散模型的生成范式，而里面最有代表性的就是Stable Diffusion，Imagen，Midjourney等。2022年初，大家把扩散模型应用到视频领域，然后就开始了扩散视频模型的大爆发吧。

视频生成技术奠基作

- VDM by Google
- Make-a-Video by Meta
- Imagen Video by Google
- VideoLDM by Nvidia（Align Your Latent paper）

在探索从text to image到text to video的路上，肯定是希望最大程度依赖pre-train好的t2i模型。video主要的问题在于它的数据是3D的，而图像的数据是2D的。这里如果直接把原来Unet的通道加一条变成3维的，估计就很难用上之前t2i的训练成果了。于是大部分的技术路线都是采用伪3D的办法，也就是（2+1）D的办法，可以理解为前面的2依然是每一帧图像的二维，而那个1是代表时间序列。每个视频都可以看成若干张图片的时序集合。这样的话，每一帧的spatial部分还是可以用依赖图像模型，而只需要重点解决时序部分，前后帧的consistency问题就好了。

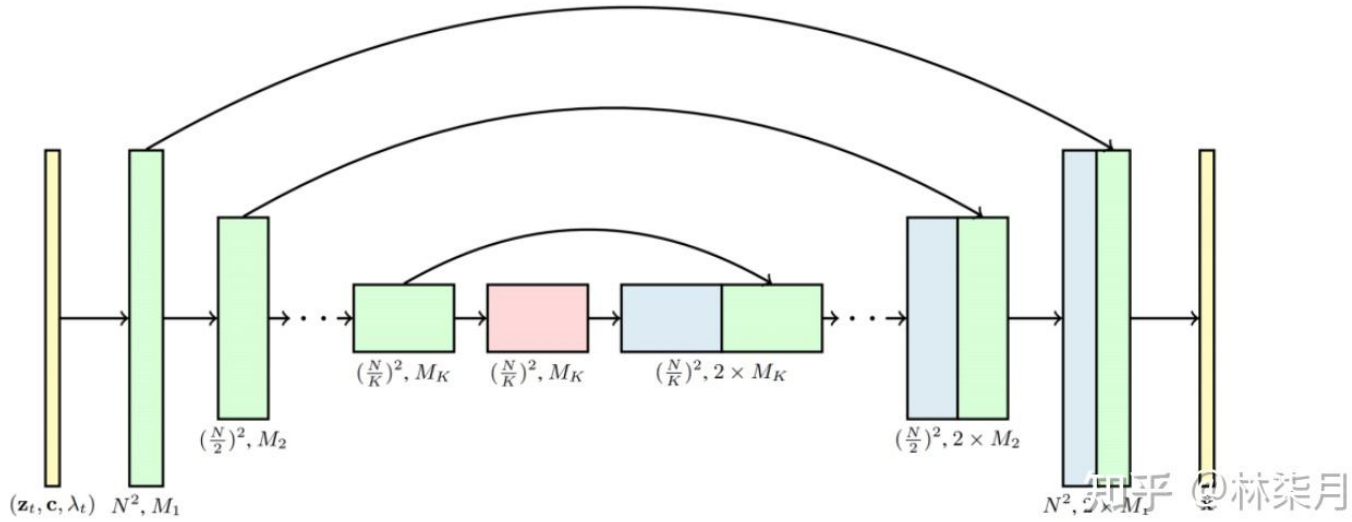


3D vs Pseudo 3D

早期这种foundation models肯定是要大公司大实验室来做的。

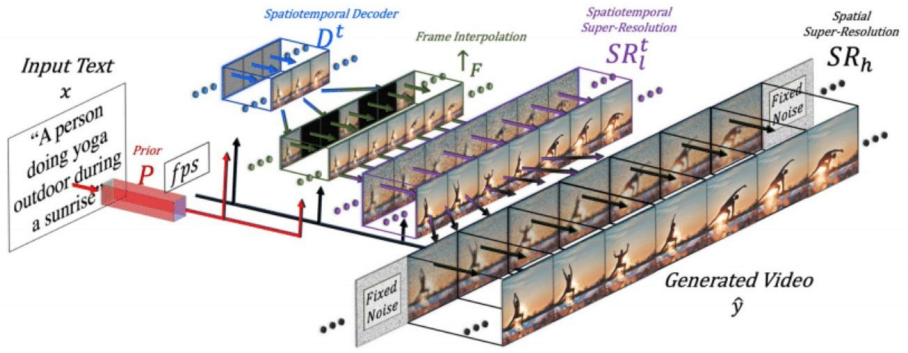
首先是google的VDM这篇paper，这应该是第一篇在视频领域的开山之作，作者也是原来做diffusion提出DDPM的大牛。它其实就是我刚刚说的采用了伪3d的办法，spatial attention直接用原来的图像模型，而额外插入了temporal attention layer。其实

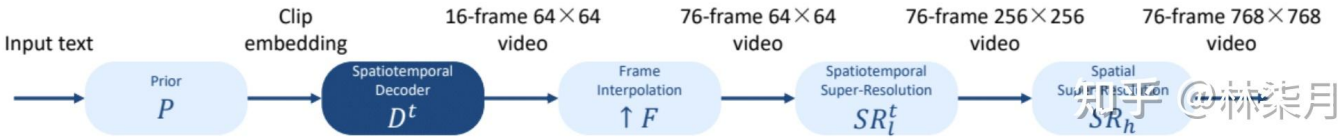
demo看多了就可以发现，视频生成最难的就是时间维度的统一性，比如前后帧连不连贯，是否smooth等，而每一帧的分辨率都是很好解决的，因为已经有很成熟的图像生成技术了。最开始这些demo的视频都只有几秒，内容也没有什么变化，基本上就是一个scene大概动一下。因为时间一长了就特别容易不连续。



from VDM

接下来是Meta做的Make a Video paper，里面用了另一种cascaded模型结构。所谓cascaded也就是采用搭积木的办法（类似Imagen），每一步实现一个goal然后串联在一起。比如这篇paper里的模型结构主要由四个主要模块组成，采用的就是先生成浓缩版本低分辨率视频（spatial temporal decoder，这个应该是核心diffusion的部分）--扩长版本低分辨率视频（interpolation）--次高分辨率视频--再到高分辨率的办法分步骤生成尽可能高质量视频。





Make a Video

Imagen Video也是follow原来Imagen的那种cascaded生成路线，模块堆的跟上面略有不同但是技术原理都是伪3D+SpatialTemperal的办法。包括Align your latent也是这个路线，只是具体的模块网络结构前后顺序不太一样，以及训练策略不同。

其他base model

这些奠基作出来之后，视频生成的路线基本上就确定了，但是模型并没有开源。后面其他有足够卡的公司/实验室也在这些基础上自己训练了各种视频大模型，以下是开源的几个：

- 阿里做的ModelScopeT2V，可以handle不同长度的帧数。Zeroscope是在它基础上在更高质量的数据集上finetune过的模型。
- 新国立的Show-1，主要argue的是better text-video alignment，使生成的视频更符合文本描述。方法是不在latent space训练而是直接pixel level。
- VideoCrafter，也是直接在SD里加temperal layer
- LaVie
- stability ai的stable video diffusion（这个是直接follow的stable diffusion了）。paper里主要argue的是scaling up训练数据集。因为视频的有标注训练资料比较少，大多都是没有caption的，那么它就弄了一个piipeline给无标注视频加caption，主要利用了各种image captioner和LLM。
- （感觉基本上都是3 stage，先把原来的t2i模型拿来用或者再train一下，然后加点temporal layer扩成伪3D的，准备训练资料。最后看是在模型架构里cascaded直接生成高分辨率，还是生成好了低分辨率版本，再在高分辨率数据上finetune整个model。）

还有闭源的：

- GenTron，W.A.L.T.，这俩都是Transformer based的模型架构

视频生成

Control

有了比较确定的视频生成技术路线之后大家也逐渐开始探索它的个性化/可控性之类的了。跟t2i领域的研究很类似，如何用更少的训练成本去加强image/video生成的可控性（比如给张草图，给关键帧，给指定姿势）或者是个性化（比如给定风格，外貌）一直是非常 的研究方向，各种paper看都看不过来。不过总体来说都是大同小异，大部分走的是ControNet/LoRA（在模型里加插件）或者Dreambooth的路线。这里tutorial提到了几个比较巧妙有代表性的。

首先是非常火的AnimateDiff，做个性化的视频生成，可以直接让一张图动起来。它的特别点在于并不是给每种风格都单独train一个video版本的LoRA或者dreambooth，而是训练了一个通用的temporal layer。生成视频的特点不在于画面怎么动（比如motion路线之类的，它就是让画面动起来而已），而是它这个可插入的temporal layer可以直接插在不同的个性化t2i model上，生成的质量就非常高。相当于这个个性化的部分是直接leverage之前t2i的成果。

Text2Video-Zero主打的是直接用图像生成模型SD来生成视频，没有finetune也不改模型结构。让SD生成一个series的frame，通过各种technique让这些frame比较相似，组合起来就是视频。

长视频生成

生成长视频最大的难题大概就是连续性。早期视频都是同一个scene动一下，只有几秒钟，但是如果生成中长视频（几分钟）的话就没法看，前后没有呼应。所以大家就通过比如利用LLM来写脚本分镜（VideoDirectorGPT，LLM-Grounded VDM等），或者给关键帧的办法让长视频更连续。

这里很喜欢的是NUWA-XL，感觉方法很巧妙，是用了recursive diffusion的办法来给中间关键帧，再用coarse-to-fine层次生成。这样就大概控制了每个时间范围内的视频语义。

多模态

除了text input还支持其他的input，比如MCDiff （sketch控制引导motion），声音input（AADiff）还有MRI信号input的那种，就是比较扯。

视频编辑

视频编辑一直是非常火的方向，因为是在本来的视频上做修改而不是从0到1生成新的，效果普遍都很好，使用价值高，图片编辑同理。不过我感觉有些视频编辑工作也可以看成给定一个reference video来增加视频生成的可控性，有一些overlap。有几个印象深刻的：

- Tune a Video，主打one-shot也就是在一个ref video上训练，学习这个video的structural info和motion。然后就可以在infer的时候换主体换风格。
- Deamix，few shot learning，直接follow的dreambooth办法
- Motion Director，可以几个ref video里面学motion
- FateZero，主要变attention maps，做了一个主体和bg分开的动作（mask）可以分别edit
- Runaway的Gen-1，支持深度图
- VideoComposer，把各种condition input（深度图，sketch，motion等）都融合起来做work
- 专门做人的：MagicAnimate，DreamPose（pose control）

- 点对点：VideoSwap等，感觉用的都类似image里dift那类的办法，擅长subject swapping
- 融合3D，用3D或者类3D表征来表示视频从而方便edit：Layer Neural Atlas（Atlas-based路线），CoDeF（用canonical image+deformation field，从而修改视频=只修改一张canonical image），DynVideo-E（直接用Nerf）
- 其他：instructVid2Vid（把instructPix2Pix的pipeline搬过来）

强烈推荐希望对此领域有更详细了解的科研同学去看原视频。

https://sites.google.com/view/showlab/tutorial