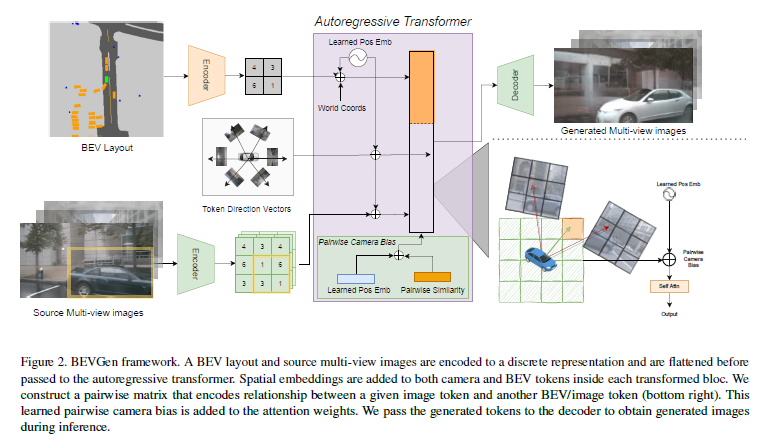
### Street-View Image Generation from a Bird's-Eye View Layout

#### 发布信息

* **发布时间**: 22 Nov 2023
* **发表位置**: Robotics and Automation Letters, 2024
* **作者**: Alexander Swerdlow, Runsheng Xu, Bolei Zhou
* **任务：自动驾驶场景数据生成-传感器数据方向**



#### 动机

从给定的BEV语义布局生成逼真街景图像的生成性BEV感知任务尚未被充分探索，这是首次尝试探索从**BEV布局去生成驾驶图像**。

#### 方法

文中提出了BEVGen，这是一种条件生成模型，可合成一组与交通场景的BEV布局相匹配的逼真且空间一致的环视图像。输入主要为两个分支，一个分支为BEV layout分支，BEV Layout通过VQ-VAE的Encoder层，得到离散化的BEV token，另一个分支为相机拍摄的多视角图像，图像通过VQ-VAE的Encoder层，得到离散化的Image token。下一阶段，将BEV token与世界坐标结合输入自回归Transformer中，将BEV token和Image token的相似性作为特征，与Image token结合，对相机之间的特征对齐和相机与BEV之间的特征进行对齐后，作为自回归Transformer的输入。其中，相似性特征是可学习的，便于生成空间一致性更好的图片。除此之外，为了实现场景中较好效果的方向对齐，将编码后的方向向量也加入到自回归Transformer中，通过Transformer后的数据经过Decoder,得到的输出是静态的多视角街景图像。这些图像是基于给定的BEV布局生成的，目的是在模拟或真实的交通场景中从不同的摄像头视角产生一致且现实感强的街景图像。

Tips:

1. 文章中提到的“source multi-view images”指的是一组从不同虚拟相机视角捕获的街景图片。这些图片不是单独的内参或外参矩阵，而是实际的图像数据，它们是从多个相机视角生成的，用以训练 BEVGen 模型，使其能够学习从 BEV 布局生成一致且逼真的街景图像。"source multi-view images"是实际的街景图片，它们与相机的内参和外参一起作为模型训练的输入，而不是仅包含参数矩阵。

2.输入输出的区别是：输入的是二维BEV布局和不同视角下的相机图像，BEVGen生成的是一组与BEV布局相匹配的现实中的空间一致的多视图街景图像。这些图像从多个虚拟相机视角生成，能够准确渲染道路和车道线，并且能够生成具有不同天气条件和一天中不同时间的交通场景图像。

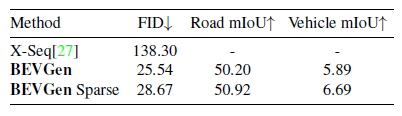
3.能够生成不同条件下的街景图像的原因不是因为使用了prompt，而是作为一个条件生成模型，能够根据输入的BEV布局和相机参数来控制生成图像的内容和视角。在训练过程中接触到了包含各种条件的街景图像，如不同的天气状况、照明条件、时间（白天或夜晚）等。因此，它能够学习到在这些不同条件下街景图像的特征，并在生成时重现这些条件。

4.输入和输出车辆不同的原因是训练过程中，BEVGen接触到了包含各种类型车辆的街景图像。这使得模型能够学习并理解不同车辆的外观特征，并在生成图像时重现这些特征。输入的BEV中并无BEVGen被训练为能够泛化到未见过的场景和条件。这意味着即使在训练数据中没有出现过特定的车辆组合，模型也能够根据其学习到的知识生成具有不同车辆的图像。

5.训练过程的输入为BEV和source multi-view Image，在训练阶段，模型使用多视图图像来学习如何从 BEV 布局生成一致的街景图像。一旦训练完成，模型就能够使用这些学到的知识来生成新的图像，而不再依赖于特定的训练图像。而推理过程的输入仅为BEV布局，相机参数在训练过程已经确定好，可以直接输出多视角街景图像。

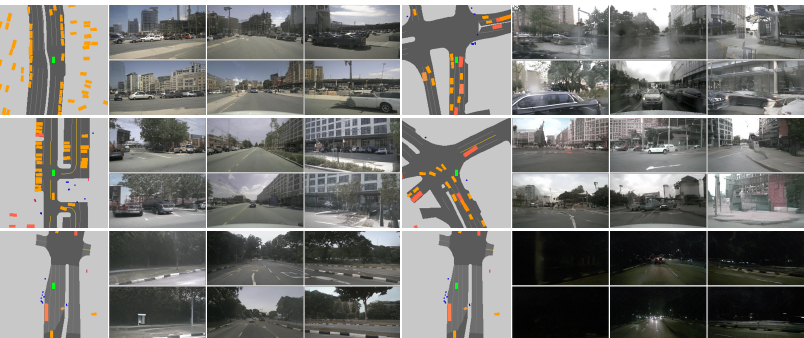
#### 实验

定量结果



定性结果

1）生成的白天与黑夜的图像



2）生成的合成图像与真实图像的比较（左为合成）



#### 结论

在这项工作中，引入了一种新的BEV生成任务，并提供了BEVGen生成模型作为该任务的解决方法。在真实世界的驾驶数据集上训练后，该模型可以从给定的汽车布局中生成空间一致的多视图图像。文中展示了它在数据扩充和模拟BEV生成中的应用。

### Panacea: Panoramic and Controllable Video Generation for Autonomous Driving

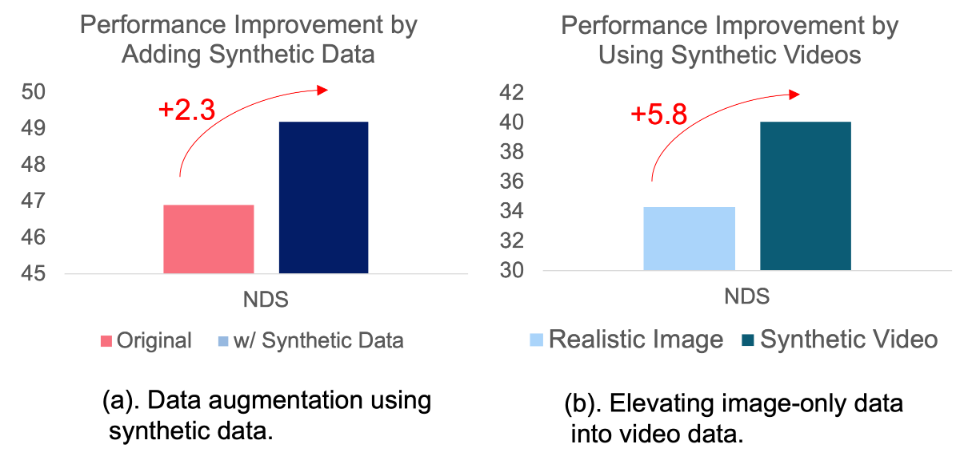
#### 发布信息

* **发布时间**: 2024
* **发表位置**: CVPR
* **作者**: University of Science and Technology of China, 2MEGVII Technology, 3Mach Drive \*Equal Contribution, †Yuqing Wen1\*†, Yucheng Zhao2\*,Yingfei Liu2\*, Fan Jia2, Yanhui Wang1, Chong Luo1, Chi Zhang3, Tiancai Wang2‡, Xiaoyan Sun1‡, Xiangyu Zhang2
* **任务：全景可控仿真数据合成**

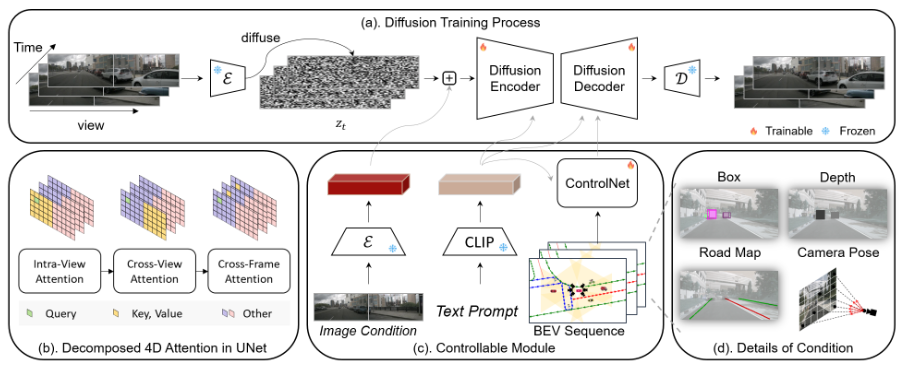
#### 动机

1.基于BEV布局序列的全景视频生成有助于建立合成视频数据集，从而增强感知任务。

2.使用条件图像和BEV布局制作全景视频可以有效地将纯图像数据集提升为视频数据集，从而推动基于视频的感知技术的发展。



#### 方法



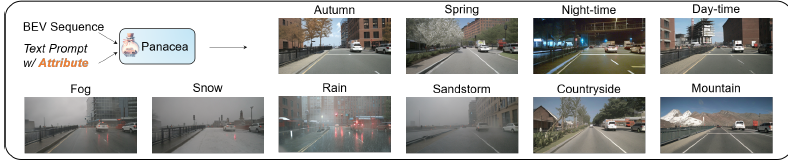
本文提出了一种叫Panacea的模型，模型的整体训练框架是一个基于强预训练的SD模型所构建的带有4D分解注意力模块的Encoder-Decoder架构。模型首先输入多视角的图像序列，通过Encoder得到噪声Zt,通过Diffusion Encoder-Decoder部分，最后解码输出全景可控图像。

生成过程中有三处创新的机制：可控模块和4D分解注意力机制和双阶段机制。

**可控模块**：

Panacea集成了两类控制信号：一类是包含文本属性的粗粒度全局控制，另一类是涉及BEV布局序列的细粒度布局控制。

粗粒度的全局控制使Panacea模型能够生成不同的多视图视频。这是通过将CLIP编码的文本prompt集成到UNet中来实现的，该方法类似于稳定扩散中使用的方法。得益于稳定扩散预训练模型，Panacea根据文本提示合成特定的驾驶场景，如图所示

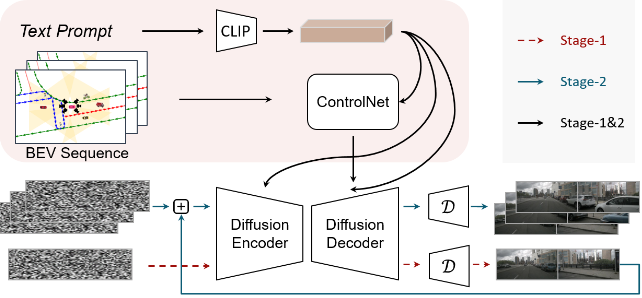


Panacea模型的细粒度布局控制有助于生成与注释一致的合成样本。我们使用BEV布局序列作为条件。具体来说，输入的BEV布局中，带有的信息有车辆的框状标注信息、灰度的深度标注信息，以及道路轮廓信息和相机的位置信息。

**4D（HWVT）分解注意力机制：**

视图视频生成可以采用4D技术来详尽地探索用于相干样本生成的多视图视频表示。然而，这种方法需要过多的内存和计算资源，甚至超过了最先进的A100GPU的能力。因此，提出了一种称为分解4D注意力的更有效的架构，将注意力分为内部视图注意力、交叉式图注意力、交叉帧注意力三部分，内部视图注意力是针对单个图片内部，交叉式图注意力针对相邻view的图片，交叉帧注意力针对相邻time的图片。

**两阶段的pipeline机制**：



在训练过程中，首先训练一组单独的权重，专门用于多视图图像生成。然后通过将条件图像与扩散输入连接起来，来训练第二阶段的视频生成权重。该条件图像仅与第一帧集成，而随后的帧采用零填充。值得注意的是，在第二阶段的训练中，使用地面实况图像而不是生成的图像作为条件。这种方法使训练过程具有与单级视频生成方案相当的效率。

在推理过程中，首先使用第一阶段的权重对多视图帧进行采样。随后生成多视图视频，该多视图视频以最初生成的帧为条件，采用第二阶段的权重。这种两阶段流水线显著提高了视觉逼真度，这归因于空间和时间合成过程的分解。不同任务及定性结果示意图如下

Tips

Two-Stage Pipeline是Panacea模型中用于增强生成质量的一种策略，它既涉及训练阶段也涉及推理阶段，它在训练阶段用于优化模型权重，在推理阶段用于生成高质量的多视图视频序列。下面分别解释两阶段流水线在训练和推理中的应用：

训练阶段:

1. **第一阶段训练**：在训练的第一阶段，模型专注于学习如何生成多视图图像。这意味着模型在这个步骤中接收文本提示和BEV布局序列作为输入，并优化用于多视图图像生成的权重。
2. **第二阶段训练**：在训练的第二阶段，模型学习如何将这些多视图图像扩展成视频序列。这一阶段中，模型会接收第一阶段生成的多视图图像作为条件输入，并训练用于视频生成的权重。

推理阶段:

1. **第一阶段推理**：在推理的第一阶段，模型使用训练好的第一阶段权重来生成多视图图像。这些图像是基于文本提示和BEV布局序列生成的。
2. **第二阶段推理**：在推理的第二阶段，模型使用第一阶段生成的多视图图像，结合后续的BEV布局，利用训练好的第二阶段权重来生成视频序列。

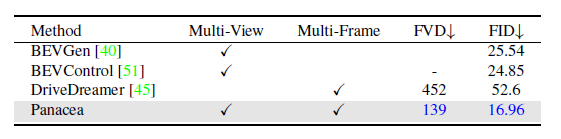
两阶段流水线的优势:

* **空间和时间合成过程的分离**：通过将图像生成和视频序列生成分开为两个阶段，模型可以更专注于每个任务，从而提高生成视频的视觉保真度和连贯性。
* **训练效率**：在第二阶段训练中，使用真实的图像而不是生成的图像作为条件输入，这有助于提高训练效率，并避免可能的累积误差。
* **生成质量**：两阶段方法能够显著提升生成视频的质量，如在文件中提到的，通过这种分解方法，FVD和FID指标都得到了显著的改善。

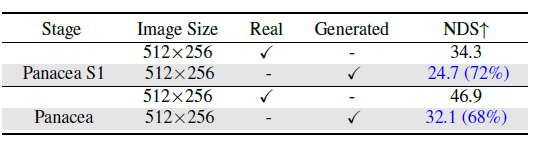
#### 实验

NDS：nuScence detection score

图片生成质量定量结果如下表

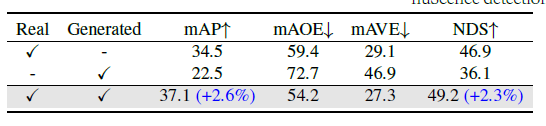


可控性能定量结果（Pannacea的相对性能）如下表：

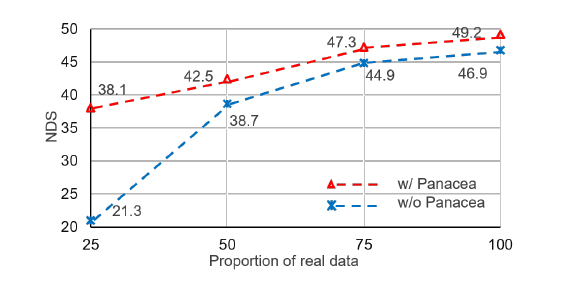


使用Stream-PETR获得的感知性能指标对控制的不稳定性进行了定量评估。首先通过Panacea生成nuSences的整个验证集。然后使用预先训练的StreamPETR模型来推导感知性能。

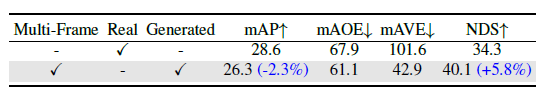
除了对验证集的评估之外，Panacea更重要的特征是它能够生成数量有限的注释训练样本。利用这一点，我们为nuScenes合成了一个新的训练数据集，名为Gen nuScene，作为StreamPETR模型的辅助训练资源。有趣的是，专门在Gen nuScenes上训练的StreamPETR模型获得了36.1%的显著NDS，与在实际nuScenes-训练集上训练的模型相比，达到了相对性能的77%，。更重要的是，将生成的数据与真实数据相结合，提高了StreamPETR模型的训练（2.3%），如下表所示



此外，下图证明了Gen-nuScenes在各种实际数据比率中始终支持Stream-PETR的性能。这些结果集体证实，我们的Panacea模型擅长生成可控的多视图视频样本，构成了自动驾驶系统的宝贵资产。



Panacea模型的另一个应用场景是它可以通过使用真实图像作为条件，将仅图像的数据集提升为视频数据集。这种提升允许应用先进的基于视频的技术。如下表所示



该过程产生了显著的改善，NDS增加了5.8个百分点。不幸的是，还注意到mAP下降了2.3个百分点，假设这是由于生成的样本与实际样本相比质量较差，以及生成的训练数据和真实验证数据之间的域不匹配。因此希望未来在提高普通样品质量方面的进展将改善这种观察到的降解。

#### 结论

Panacea是一款尖端的生成器，专门用于为驾驶场景生成可控的全景视频。在Panacea中，我们结合了一个分解的4D注意力模块，以确保时间和跨视图的一致性，促进生成逼真的多视图图像和视频。同时，采用两阶段训练策略，进一步提高生成质量。值得注意的是，Panacea擅长处理各种控制信号，以生成具有精确注释的视频。通过广泛的实验，Panacea展示了其制作高质量、注释良好的全景驾驶场景视频的能力。这些视频是无价的，不仅在生成汽车感知方面发挥作用，而且在现实世界的驾驶模拟中也有前景。

### ADriver-I: A GeneralWorld Model for Autonomous Driving

#### 发布信息

* **发布时间**: 2024
* **发表位置**: CVPR
* **作者**: Fan Jia1⋆ Weixin Mao2⋆† Yingfei Liu1 Yucheng Zhao1 Yuqing Wen3†

Chi Zhang4 Xiangyu Zhang1 Tiancai Wang1‡

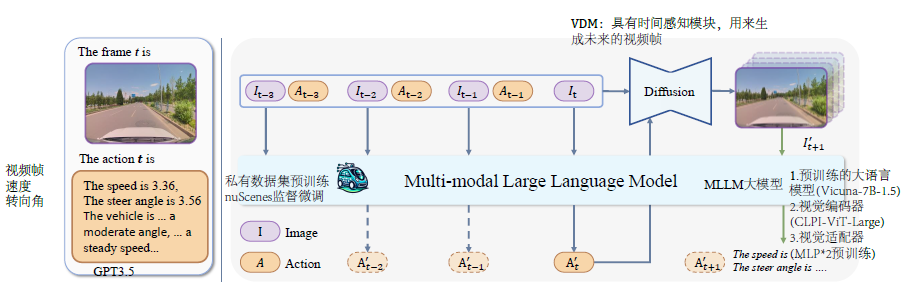
1MEGVII Technology 2Waseda University

3University of Science and Technology of China 4Mach Drive

* 任务：场景预测

#### 方法

提出了一种MLLM和扩散模型结合的模型，成为ADriver-I。它以视觉-动作对为输入，逐步预测当前帧的控制信号。所生成的控制信号与历史的视觉-动作对一起，进一步调节去预测未来帧。利用预测的下一帧，ADriver-I执行进一步的控制信号预测。这样的过程可以无限重复，ADriver-I实现了自我创造世界中的自动驾驶。



具体方法是输入四帧的视觉-动作对，（It−1，At−1）...，（It−3，At−3）用作MLLM和VDM的输入。在训练过程中，MLLM以自回归的方式输出低级控制信号集{At−2，At−1，At}。之后输出的控制信号At被进一步用作VDM的提示，预测接下来的四帧{It+1，…，It+4}。

在推断过程中，MLLM单步直接输出当前帧At的控制信号。VDM遵循相同的过程并预测未来帧——生成的下一帧It+1用作下一时间的“当前帧”，并进一步输入到MLLM产生动作At+1。上述步骤可以循环重复。通过这种方式，它实现了由自己生成世界中的无限驾驶。

#### 结果

为了定量评估预测的未来帧的质量，计算FID和FVD指标如下表所示。

