I. 方法

本章给出所提 **SiT-PVG** 的完整方法。按顺序介绍：预备知识（§3.1），2D→4D 语义蒸馏（§3.2），双线索动态掩码（§3.3），语义驱动的时间约束（§3.4），时间一致性增强（§3.5），以及优化与实现细节（§3.6）。

A. 预备知识：3DGS 与 PVG

3D Gaussian Splatting（3DGS） 以一组各向异性的高斯基元集合显式建模场景。每个基元包含空间中心、各向异性形状与朝向、不透明度与外观参数；通过可微光栅化与按深度排序的透明度融合实现高效渲染与快速收敛。高斯基元的数学表达式为：

$$G_i(\boldsymbol{x}) = \exp\Big(-\tfrac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i) \Big). \tag{1}$$

其中，$\boldsymbol{\mu}_i$ 为三维位置；$\boldsymbol{\Sigma}_i$ 为协方差矩阵，描述形状与朝向，通常由旋转矩阵 $\mathbf{R}$ 与尺度矩阵 $\mathbf{S}$ 参数化（如 $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{S}\mathbf{S}^\top\mathbf{R}^\top$）。将三维高斯投影到像平面得到 2D 高斯，其投影协方差由

$$\Sigma' = JW\,\Sigma\,W^\top J^\top \tag{2}$$

给出，其中 W 是世界到相机的外参变换（SE(3)），J 为透视投影的雅可比近似。像素颜色采用按深度排序的 α-融合：

$$C = \sum_{i=1}^{N} T_i\,\alpha_i\,c_i, \qquad T_i = \prod_{j<i}(1-\alpha_j), \tag{3}$$

其中 α_i 由点元不透明度与其投影协方差在该像素的覆盖贡献共同决定，c_i 为外观（如球谐系数着色）。上述表示配合基于瓦片的可微光栅化，使 3DGS 在静态场景中实现实时渲染和快速收敛。

3DGS 在建模上默认静态：点元参数随时间不变，难以直接刻画道路场景中普遍存在的时变要素（车辆、行人等）。为此，**Periodic Vibration Gaussians（PVG）** 在 3DGS 的最小改动上引入时间参数化：令点元的空间位置与不透明度随时间围绕“寿命峰值” τ 作可微振荡与衰减。具体地，对每个点元引入周期长度 l、速度方向/幅度 v、寿命尺度 β，定义

$$\tilde{\mu}(t) = \mu + \frac{l}{2\pi}\sin\Big(2\pi\frac{t-\tau}{l}\Big)\,v, \qquad \tilde{o}(t) = o\cdot\exp\Big(-\tfrac{1}{2}\,(t-\tau)^2\,\beta^{-2}\Big). \tag{4}$$

此时点元在时刻 t 的状态为 $H(t) = \{\tilde{\mu}(t), q, s, \tilde{o}(t), c\}$，整幅图像按

$$\hat{I}_t = \mathrm{Render}\big(\{H_i(t)\}_{i=1}^{N}; E_t, I_t\big) \tag{5}$$

渲染。为衡量点元“静态度”，定义

$$\rho = \beta / l, \tag{6}$$

ρ 越大表示寿命相对周期更长、越趋近静止；当 $v=0$ 且 $\rho \to \infty$ 时，PVG 退化回标准 3DGS。由此，静态/动态以统一参数化出现，仅通过 $\{v, \beta, l, \tau\}$ 的取值加以区分。PVG 以最小改动继承了 3DGS 的高效与可扩展性，同时补足了动态建模与可编辑性，是本文面向道路环境的更合适表征选择。

B. 2D-4D 语义蒸馏

为解决仅凭重建损失难以稳定区分“相机运动”与“真实世界运动”的问题，并让模型更好地理解场景语义，我们将 2D 基础模型的稠密语义特征迁移至 4D 高斯表征，使每个高斯点元学习到连续可度量的语义向量，便于在后续构造语义先验与动静掩码。该范式已在 3DGS/4DGS 框架中验证有效，本文将其无缝引入 PVG 框架。

我们采用 LSeg 作为教师模型。其像素特征与 CLIP 文本空间对齐，能够提供连续可度量、开放词汇的语义向量。对时刻 t 的真帧 I_t 提取像素对齐的教师特征图

$$F_t = \mathrm{LSeg}(I_t).$$

对学生端，我们为每个高斯基元赋予可学习语义向量 $f_{\mathrm{sem},i}$。与 RGB 渲染完全一致，像素 $\mathbf{p}$ 处的学生语义由可见性 α 合成权重进行加权聚合；记 $\mathcal{V}(\mathbf{p},t)$ 为按深度排序的可见点元集合、$w_i(\mathbf{p},t)$ 为对应的 α-合成权重，则

$$F_s(\mathbf{p},t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,f_{\mathrm{sem},i}. \tag{7}$$

为与教师通道数对齐，我们使用轻量线性头（1×1 卷积/全连接） $U(\cdot)$ 将学生输出映射到教师维度：

$$\tilde{F}_s(\mathbf{p},t) = U\big(F_s(\mathbf{p},t)\big).$$

我们采用像素级 L_1 蒸馏使学生贴近教师，构建蒸馏损失

$$\mathcal{L}_{\mathrm{SD}} = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}\Big\|\tilde{F}_s(\mathbf{p},t) - F_t(\mathbf{p})\Big\|_1,$$

其中 Ω 为当前分辨率下的像素集合。训练收敛后，参与该像素合成的高斯会将教师语义“写入”其 f_{sem}，形成可在时序上随 PVG 形变进行搬运的 4D 语义表征。与语义先验及双线索动静掩码的融合将于§3.3 详述。

C. 双线索动态掩码（Dual-Evidence Motion Mask)

在§3.2 中，我们已获得教师的像素语义特征 F_t 与学生侧的语义渲染 $\tilde{F}_s$。本节在此基础上，基于两类证据构建帧级动静掩码：其一为教师—学生特征差异，其二为语义先验。前者刻画像素处的语义不一致性，后者提供先验上应静止的区域指示。两者经轻量融合得到稳定的静态掩码 M_{stat} 与动态掩码 M_{dyn}，用于路由后续时间约束。

a) (a) 教师—学生特征差异: 当像素对应静态背景时，多帧观察通常指向同一世界点，学生的语义渲染应与教师特征一致；而当像素位于动态目标或遮挡边界时，学生渲染更易与教师产生偏差。基于此，定义像素级余弦不相似度：

$$D(\mathbf{p},t) = 1 - \cos\big(\tilde{F}_s(\mathbf{p},t),\, F_t(\mathbf{p})\big). \tag{8}$$

D 越大，表示该像素更可能存在真实运动或配准不稳定。

b) (b) 语义先验: 利用 LSeg 的像素特征与文本原型得到类别分数 $\{S_k(\mathbf{p},t)\}$，将静态倾向类（如 road/building/sky）累加为软静态先验：

$$M_{\text{sem}}(\mathbf{p},t) = \sum_{k\in\mathcal{C}_{\text{stat}}} S_k(\mathbf{p},t). \tag{9}$$

c) (c) 轻量融合与阈值化: 我们采用简单的的逻辑回归层融合两条证据并输出静态概率 $\delta \in (0,1)$：

$$\delta(\mathbf{p},t) = \sigma\big(a\cdot(1-D(\mathbf{p},t)) + b\cdot M_{\text{sem}}(\mathbf{p},t) + c\big), \tag{10}$$

其中 $\sigma(\cdot)$ 为 Sigmoid，$a,b,c \in \mathbb{R}$ 为可学习标量（初始化 a=1, b=1, c=0），与主网络共同训练。训练期间，我们将 δ 作为软静态权重直接参与后续时间约束（§3.4）的加权并采用固定阈值将其二值化：

$$M_{\text{stat}}(\mathbf{p},t) = \mathbf{1}\big(\delta(\mathbf{p},t) > \tau_s\big), \qquad \tau_s = 0.5, \tag{11}$$

并令 $M_{\text{dyn}}(\mathbf{p},t) = 1 - M_{\text{stat}}(\mathbf{p},t)$。此外，我们在实现中采用“保守融合”策略，即静态取并，动态取交，以进一步减小误差。

D. 语义驱动的时间约束

为降低静态背景的伪运动并维持动态目标的时间连贯性，在不增加体渲复杂度的前提下，将“静区应近零速度、动区允许合理运动、跨时间保持稳定”的先验以可微方式注入模型。核心思路是：利用§3.3 的像素级静态概率 $\delta(\mathbf{p},t)$ 及其回投得到的点元级静态概率 w_i^{stat}，在参数层约束点元的速度与寿命（SVC/SLP）。

a) 语义速度约束（SVC）: 以点元语义向量 $f_{\text{sem},i}$ 产生速度门控并直接作用于 PVG 的速度基向量：

$$g_i = \sigma(\mathbf{w}_g^\top f_{\text{sem},i} + b_g) \in (0,1), \qquad \boldsymbol{v}_i^{\text{eff}} = g_i\,\boldsymbol{v}_i, \tag{12}$$

随后以 $\boldsymbol{v}_i^{\text{eff}}$ 替代 $\boldsymbol{v}_i$ 更新轨迹 $\boldsymbol{\mu}_i(t)$（其余渲染流程不变）。为度量残余运动，采用对称时间步 Δ 的投影位移并按与 RGB 相同的 α-合成权重 $w_i(\mathbf{p},t)$ 聚合得到像素速度图：

$$V(\mathbf{p},t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,\big\|\Pi\big(\boldsymbol{\mu}_i(t+\Delta)\big) - \Pi\big(\boldsymbol{\mu}_i(t-\Delta)\big)\big\|_1. \tag{13}$$

训练时不对 δ 二值化，而是将其作为软静态权重抑制静区速度：

$$\mathcal{L}_v = \text{mean}_{\mathbf{p}\in\Omega}\big(\delta(\mathbf{p},t)\cdot V(\mathbf{p},t)\big). \tag{14}$$

b) 静态寿命先验（SLP）: 仅限制瞬时速度仍可能在长序列中积累缓慢漂移。基于 PVG 的寿命尺度 β_i 与周期长度 l_i 定义静态度

$$\rho_i = \beta_i / l_i, \tag{15}$$

并用 w_i^{stat} 对静态倾向更高的点元施加下界约束：

$$\mathcal{L}_\rho = \sum_i w_i^{\text{stat}}\,\max\big(0,\ \rho^\star - \rho_i\big), \tag{16}$$

其中 $\rho^\star > 0$ 为静态度下界超参数，控制“长寿命、低频摆动”的偏好强度；ρ_i 越大表示该点更接近真正静止的表面。

E. 优化（Optimization)

训练目标需要在“外观/几何重建、语义对齐、时间稳定”之间取得平衡。经验表明，若在几何尚未收敛时过早强化时间约束，容易引发训练抖动或把真实运动过度抑制。为此，我们采用温启动策略：先以重建与语义蒸馏稳定外观与语义，再逐步注入时间约束的影响。本文中的重建项 $\mathcal{L}_{\text{rgb}}$ 由像素级 L_1 与 SSIM 的加权组合构成，用于驱动几何与外观的基础收敛而不额外引入假设。

总体目标定义为

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{17}$$

其中 $\mathcal{L}_{\text{SD}}$ 为§3.2 的语义蒸馏损失，$\mathcal{L}_v$ 与 $\mathcal{L}_\rho$ 分别对应§3.4 的语义速度约束与静态寿命先验，$\mathcal{L}_{\text{reg}}$ 为轻量正则。为避免早期过度抑制动态与造成不稳定，权重采用分阶段调度：λ_{SD}=1.0（常数）；λ_v 在前 5k 次迭代由 0 线性升至 0.5；$\rho^\star$ 在前 15k 次迭代由 1.0 线性升至 1.5；$\lambda_\rho \equiv 0.15$（或在前 2k 次迭代从 0 线性预热到 0.15）。

为抑制语义向量与速度基过大波动，我们引入 ℓ_2 正则：

$$\mathcal{L}_{\text{reg}} = \lambda_f \sum_i \|f_{\text{sem},i}\|_2^2 + \lambda_v^{\ell_2} \sum_i \|\boldsymbol{v}_i\|_2^2, \tag{18}$$

默认 λ_f=1×10^{-4}，$\lambda_v^{\ell_2}$=5×10^{-5}。时间相关参数采用中性初始化以避免早期过抑制：$\boldsymbol{v}_i$=**0**、l_i=1、β_i=1；语义门控与动静融合器的线性参数初始化为零（$g_i \approx 0.5,\ \delta \approx 0.5$），随后与主网络端到端共同更新。