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 $f_{\text{sem},i}$。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

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

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

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

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）

仅依赖重建误差或单一线索容易把“相机运动”误判为“世界运动”，从而在道路背景处产生伪动。为此，本文在每一帧上构造由两类证据共同约束的动静掩码：教师–学生特征差异与语义/实例先验。前者反映“像素处是否存在语义不一致或真实运动”，后者提供“哪些区域先

验上应静止”的弱监督。二者经简单的学习型融合与时空投票得到稳定的静态掩码 M_{stat} 与动态掩码 M_{dyn}，用于路由后续的时间约束（§3.4）。

a) (a) 特征差异线索: 设§3.2 中的教师特征为 F_t，学生对齐后的特征为 $\tilde{F}_s$。为避免与蒸馏目标相互牵制，对学生分支停止梯度，定义像素级余弦不相似度：

$$D(\mathbf{p},t)=\frac{1}{2}\Big(1-\cos\big(\mathrm{sg}[\tilde{F}_s(\mathbf{p},t)],\ \tfrac{F_t(\mathbf{p})}{\|F_t(\mathbf{p})\|_2}\big)\Big)\in[0,1],\quad (8)$$

其中 sg[·] 表示 stop-gradient。D 越大，表示教师与学生在该像素越不一致，更可能为动态或遮挡/配准困难区域。

b) (b) 语义/实例先验: 以 LSeg 的像素特征与文本原型获得类别分数 $\{S_k(\mathbf{p},t)\}$（或直接使用 LSeg 的分割输出），将静态倾向类（如 road/building/sky 等）累加得到软静态先验：

$$M_{\text{sem}}(\mathbf{p},t)=\sum_{k\in\mathcal{C}_{\text{stat}}}S_k(\mathbf{p},t)\in[0,1].\quad (9)$$

可选地，使用 SAM/实例分割在边界处细化先验，但不改变本文的主体流程。

c) (c) 学习型融合与自适应阈值: 为避免手工权重，我们用一条极简的逻辑回归把两条证据融合为“静态性分数”：

$$\delta(\mathbf{p},t)=\sigma\big(a\cdot(1-D(\mathbf{p},t))+b\cdot M_{\text{sem}}(\mathbf{p},t)+c\big)\in(0,1),\quad (10)$$

其中 $\sigma(\cdot)$ 为 sigmoid，a,b,c 为可学习标量。再用帧内自适应阈值得到学习型静态掩码（η 为分位数，如 0.6）：

$$\varepsilon_t=\mathrm{Percentile}\big(\{\delta(\mathbf{p},t)\}_{\mathbf{p}\in\Omega},\,\eta\big),\qquad M_{\text{learn}}(\mathbf{p},t)=\mathbf{1}\big(\delta(\mathbf{p},t)>\varepsilon_t\big).\quad (11)$$

最后采用“保守融合”（静态取交、动态取并），得到初始动静掩码：

$$M^0_{\text{stat}}=M_{\text{learn}}\wedge\mathbf{1}(M_{\text{sem}}>\tau_{\text{sem}}),\qquad M^0_{\text{dyn}}=1-M^0_{\text{stat}},\quad (12)$$

其中 $\tau_{\text{sem}}\in(0,1)$ 为静态先验的置信阈值（如 0.5）。

d) (d) 时空一致性投票: 为抑制瞬时噪声与遮挡，我们将相邻时刻/视角的静态掩码回投到当前帧并做多数投票。记邻域 $\mathcal{N}_t$，几何回投算子为 $W_{u\to t}(\cdot)$（由相机标定与可见性计算），则

$$\bar{M}_{\text{stat}}(\mathbf{p},t)=\mathbf{1}\left(\frac{M^0_{\text{stat}}(\mathbf{p},t)+\sum_{u\in\mathcal{N}_t}W_{u\to t}\big(M^0_{\text{stat}}(\cdot,u)\big)}{1+|\mathcal{N}_t|}>\tau_{\text{vote}}\right)\quad (13)$$

其中 $\tau_{\text{vote}}\in(0,1)$（如 0.5）。我们将 $\bar{M}_{\text{stat}}$ 作为最终静态掩码，$M_{\text{stat}}=\bar{M}_{\text{stat}}$，并令 $M_{\text{dyn}}=1-M_{\text{stat}}$。实践中可对 $\bar{M}_{\text{stat}}$ 做轻量形态学开闭运算以去除孤立噪点。

e) (e) 从像素到点元的静态概率: 后续的时间约束（§3.4）需在点元层面使用静态概率。将像素掩码按可见性聚合到高斯点元 g_i，得到

$$w_i^{\text{stat}}=\frac{\sum_t\sum_{\mathbf{p}\in\Pi_t(i)}w_i(\mathbf{p},t)\,M_{\text{stat}}(\mathbf{p},t)}{\sum_t\sum_{\mathbf{p}\in\Pi_t(i)}w_i(\mathbf{p},t)+\varepsilon},\quad (14)$$

其中 $\Pi_t(i)$ 为点元 i 在时刻 t 的投影支持像素集合，ε 为数值稳定项。w_i^{stat} 将直接用于§3.4 的速度门控与寿命先验的加权。

f) (f) 讨论: D 与 M_{sem} 分别从“数据一致性”和“先验语义”两侧提供互补证据：D 对真实运动与匹配困难敏感，M_{sem} 在大尺度静态背景上稳定。式 (10)–(12) 用极简的可学习融合与自适应阈值避免手工调参，式 (13) 的时空投票进一步提升掩码稳健性。由于 (8) 对学生分支采用 stop-gradient，动静掩码的构建不会与蒸馏目标相互干扰。

D. 双线索动态掩码（Dual-Evidence Motion Mask）

在§3.2 中我们已得到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，给出一套无需额外训练、直接可复现的动静掩码构建方法：以**教师–学生特征差异**作为“数据一致性”证据，以**语义先验**作为“类别常识”证据，使用固定规则融合得到静态掩码 M_{stat} 与动态掩码 M_{dyn}，用于路由§3.4 的时间约束。

a) (a) 教师–学生特征差异（数据一致性）: 为避免与蒸馏目标相互牵制，对学生分支停止梯度；定义像素级余弦不相似度（差异越大越可能为动态或匹配困难）：

$$D(\mathbf{p},t)=1-\cos\big(\tilde{F}_s(\mathbf{p},t),\,F_t(\mathbf{p})\big)=1-\frac{\tilde{F}_s(\mathbf{p},t)^\top F_t(\mathbf{p})}{\big\|\tilde{F}_s(\mathbf{p},t)\big\|_2\,\big\|F_t(\mathbf{p})\big\|_2}.\quad (15)$$

符号说明（简要）：

- $\mathbf{p}$：像素坐标（二维图像平面上的位置）。
- t：时间步或帧索引（对应真实帧 I_t）。
- $F_t(\mathbf{p})$：**教师**LSeg在真实帧 I_t 上对像素 $\mathbf{p}$ 提取的语义特征向量（未训练、前向即可得到）。
- $\tilde{F}_s(\mathbf{p},t)$：**学生**侧由4D高斯语义向量经可见性加权渲染并用线性头对齐后的像素特征（与RGB同一条光栅化管线得到）。
- $\cos(\cdot,\cdot)$：向量的余弦相似度，右式给出其标准定义（点积除以范数乘积）。

直觉： $D(\mathbf{p},t)$ 度量“学生渲出的语义”与“教师在真实图像上的语义”是否一致；值越大表示越不一致，常出现在动态物体、遮挡边界或当前尚未学稳的区域。在实践中可把 D 视作**动态/不稳定倾向**，配合语义先验与时空投票用于构建动静掩码。

b) (b) 语义先验（类别常识）： 利用 LSeg 的像素语义分数 $\{S_k(\mathbf{p},t)\}$（或其分割输出），将“静态倾向类”（如 road/building/sky 等）累加为软静态先验：

$$M_{\text{sem}}(\mathbf{p},t)=\sum_{k\in\mathcal{C}_{\text{stat}}} S_k(\mathbf{p},t)\in[0,1]. \tag{16}$$

c) (c) 固定规则融合与阈值化: 我们以固定权重将两条证据融合为“静态置信分数”，再一次阈值化得到掩码，无需引入额外的学习或投票过程：

$$S_{\text{stat}}(\mathbf{p},t)=(1-\lambda)\,M_{\text{sem}}(\mathbf{p},t)+\lambda\,\big(1-D(\mathbf{p},t)\big),\qquad \lambda\in[0,1], \tag{17}$$

$$M_{\text{stat}}(\mathbf{p},t)=\mathbf{1}\big(S_{\text{stat}}(\mathbf{p},t)\geq\tau_{\text{stat}}\big),\qquad M_{\text{dyn}}(\mathbf{p},t)=1-M_{\text{stat}}(\mathbf{p},t), \tag{18}$$

其中 λ 为两证据的平衡系数，τ_{stat} 为静态阈值（默认 λ=0.5, τ_{stat}=0.5 即可作为稳健起点）。必要时可对 M_{stat} 进行轻量的形态学开闭运算以消除孤立噪点与小孔洞。

d) (d) 从像素到点元的静态概率（供§3.4 使用）: 为在点元层面路由时间约束，将像素掩码按可见性聚合为高斯点元的静态概率：

$$w_i^{\text{stat}}=\frac{\sum_t\sum_{\mathbf{p}\in\Pi_t(i)} w_i(\mathbf{p},t)\,M_{\text{stat}}(\mathbf{p},t)}{\sum_t\sum_{\mathbf{p}\in\Pi_t(i)} w_i(\mathbf{p},t)+\varepsilon}, \tag{19}$$

其中 $\Pi_t(i)$ 为点元 i 在时刻 t 的投影支持像素集合，$w_i(\mathbf{p},t)$ 为渲染中的可见性权重，ε 为数值稳定项。

e) 直觉解释: 可以把 LSeg 的像素特征视为“老师在真图上的语义判断”，把我们用 4D 高斯渲染出的语义特征视为“学生在同一视角下的理解”。当像素对应静态背景（路、墙、楼），多帧/多视角看到的是同一个世界点，学生最终会与老师高度一致，D 很小；当像素落在动态对象或强遮挡处，负责它的高斯集合在时间上不断变化，学生难以在所有帧都与老师一致，D 会偏大。与此同时，语义先验 M_{sem} 提醒模型：路/楼/天更倾向静止、人/车更倾向运动。式 (17) 用一个固定、可解释的线性规则把两条直觉合到一起：M_{sem} 越高、D 越小，就越静；反之越动。这样得到的 M_{stat}/M_{dyn} 既简单稳健，又便于直接路由§3.4 的速度门控与寿命先验。

f) (c) 学习型融合与自适应阈值（简化版）: 我们用一条线性加权把两条证据合成为“静态分数”，去掉 sigmoid 和偏置，仅保留最直观的关系：语义先验越大、差异越小，越可能静态。

$$s(\mathbf{p},t)=\alpha\,\big(1-D(\mathbf{p},t)\big)+\beta\,M_{\text{sem}}(\mathbf{p},t), \tag{20}$$

其中 $\alpha,\beta\geq 0$ 为可学习或手动设置的权重（默认可取 $\alpha=\beta=1$）。

为适应不同帧的分布差异，采用帧内分位数做自适应阈值（$\eta\in(0,1)$，如 0.6）：

$$\varepsilon_t=\text{Percentile}(\{s(\mathbf{p},t)\}_{\mathbf{p}\in\Omega},\,\eta),\qquad M_{\text{learn}}(\mathbf{p},t)=\mathbf{1}\big(s(\mathbf{p},t)>\varepsilon_t\big). \tag{21}$$

最后，为减少误检，采用“保守融合”（静态取交、动态取并）：

$$M^0_{\text{stat}}(\mathbf{p},t)=M_{\text{learn}}(\mathbf{p},t)\wedge\mathbf{1}\big(M_{\text{sem}}(\mathbf{p},t)>\tau_{\text{sem}}\big),\qquad M^0_{\text{dyn}}=1-M^0_{\text{st}} \tag{22}$$

其中 $\tau_{\text{sem}}\in(0,1)$ 为语义先验的置信阈值（如 0.5）。

说明与直觉： (1) s 同时考虑了“学生与教师是否一致”（$1-D$）和“先验上是否应静态”（M_{sem}）；(2) 分位数阈值让各帧自动适配难度，无需手工定阈；(3) 保守融合保证只有当学习结果与先验同时支持静态时才标为静态，从而降低动态区域被误判为静态的风险。

g) (c) 轻量融合与阈值化（确定实现）: 我们采用一个两输入一输出的逻辑回归层（参数在全图共享）来融合两条证据，并输出静态概率 $\delta\in(0,1)$：

$$\delta(\mathbf{p},t)=\sigma\big(a\cdot(1-D(\mathbf{p},t))+b\cdot M_{\text{sem}}(\mathbf{p},t)+c\big), \tag{23}$$

其中 $\sigma(\cdot)$ 为 Sigmoid，$a,b,c\in\mathbb{R}$ 为**可学习的标量**（初始化 a=1, b=1, c=0），使用与主网络相同的优化器（如 Adam）在端到端训练中共同更新。训练阶段我们将 δ 作为软静态权重直接参与后续时间约束（§3.4）的加权；推理阶段采用固定阈值将其二值化：

$$M_{\text{stat}}(\mathbf{p},t)=\mathbf{1}\big(\delta(\mathbf{p},t)>\tau_s\big),\qquad \tau_s=0.5, \tag{24}$$

并令 $M_{\text{dyn}}(\mathbf{p},t)=1-M_{\text{stat}}(\mathbf{p},t)$。此外，为进一步降低误检，我们在实现中采用保守融合策略：静态判定需同时得到语义先验的支持（即与 M_{sem} 一致），动态判定则以并集为准；为简洁起见，此处不展开额外公式。

E. 语义驱动的时间约束与一致性增强（Semantic Temporal Constraints & Consistency)

本节将原§3.4 与§3.5 合并，围绕“静区应近零速度、动区允许合理运动、跨时间保持稳定”提出一组轻量且可解释的约束。核心思路是：利用§3.3 得到的像素级静态概率 $\delta(\mathbf{p},t)$ 与点元级静态概率 w_i^{stat}（由 δ 回投聚合而来），在 参数层 控制点元速度与寿命（SVC/SLP），并在 时序层 约束几何闭环与实例语义一致性（TCC/ICC）。所有项与 3DGS/4DGS 的渲染路径保持一致，不引入体积积分等昂贵算子。

a) (a) 语义速度约束（SVC）： 为避免静态背景出现抖动/漂移，我们用点元语义向量 $f_{\mathrm{sem},i}$ 产生速度门控，直接作用于 PVG 的速度基（§3.1）：

$$g_i = \sigma\big(\mathbf{w}_g^\top f_{\mathrm{sem},i} + b_g\big) \in (0,1), \qquad v_i^{\mathrm{eff}} = g_i\, v_i, \tag{25}$$

并以 v_i^{eff} 替代 v_i 带入点元轨迹 $\mu_i(t)$ 的计算（其余渲染管线保持不变）。为构造像素级的“残余速度”观测，我们以对称时间步 Δ 在像素平面上做可见性加权的位移累计：

$$V(\mathbf{p},t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,\Big\|\Pi\big(\mu_i(t+\Delta)\big) - \Pi\big(\mu_i(t-\Delta)\big)\Big\|_1, \tag{26}$$

其中 $w_i(\mathbf{p},t)$ 为与 RGB 相同的 α-合成权重。训练期间不做二值化，直接用 $\delta(\mathbf{p},t)$ 作为软静态权重抑制静区残余速度：

$$\mathcal{L}_v = \mathrm{mean}_{\mathbf{p}\in\Omega}\big(\delta(\mathbf{p},t)\cdot V(\mathbf{p},t)\big). \tag{27}$$

实现上 Δ=1（帧）即可；$\mathbf{w}_g, b_g$ 与主网络共同训练，初始可设为 $\mathbf{0}$ 与 0 以避免早期过抑制。

b) (b) 静态寿命先验（SLP）： 为提升长期稳定性，我们对静态点元的“静态度”比值 $\rho_i = \beta_i/l_i$（§3.1）施加下界约束，按可见性把像素级静态概率回投为点元级权重 w_i^{stat}（见§3.3）：

$$\mathcal{L}_\rho = \sum_i w_i^{\mathrm{stat}}\,\max\big(0,\ \rho^\star - \rho_i\big), \tag{28}$$

其中 $\rho^\star$ 训练中从 1.0 线性升至 1.5（或按数据集微调）。SVC 抑制瞬时伪动，SLP 提升长期稳态，二者互补。

c) (c) 时间闭环一致性（TCC）： 记 $\Phi_i(\cdot; t_1\to t_2)$ 为依据 PVG 时间参数将点元 i 从 t_1 映射到 t_2 的算子。为抑制随时间积累的漂移，我们对相邻时刻 $t\pm\Delta$ 施加闭环约束，仅在可见且遮挡轻微的点元上计入：

$$\mathcal{L}_{\mathrm{TCC}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \big\|\Phi_i^{-1}\big(\Phi_i(\mu_i;\, t-\Delta\to t+\Delta);\ t+\Delta\to t-\Delta\big) - \mu_i(t-\Delta)\big\|, \tag{29}$$

其中 $\mathcal{I}$ 由可见性阈值（如平均 α 覆盖）筛选得到。可将 L_1 替换为 Huber 提高遮挡边界的鲁棒性。

d) (d) 实例一致性增强（ICC，中心式）： 为稳定前景目标在跨时间的语义表征，我们以实例为单位维护语义原型 p_k（由同一实例跨时间的点元语义向量指数滑动平均得到），并在其邻域内使用中心损失：

$$\mathcal{L}_{\mathrm{ICC}} = \frac{1}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k}\big\|f_{\mathrm{sem},i} - p_k\big\|_2^2, \qquad p_k \leftarrow \eta\, p_k + (1-\eta)\,\frac{1}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k} f_{\mathrm{sem},i}, \tag{30}$$

其中 $\eta\in[0.7, 0.95]$ 控制原型更新速率，$\mathcal{S}_k$ 由实例分割（如 SAM）在时间域做轻量关联得到。若需更强判别，可替换为基于余弦相似度的 InfoNCE 形式，但我们默认采用中心式以简洁稳健。

e) (e) 总结与使用方式： 训练时，不对 δ 二值化：SVC 的零速抑制（式 (101)）直接使用 δ 的软权重；SLP 的点元权重 w_i^{stat} 由 δ 回投聚合得到；TCC/ICC 仅在可见性良好的样本上计算。上述四项与重建项、语义蒸馏项一起优化，可写为

$$\mathcal{L} = \mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{SD}}\mathcal{L}_{\mathrm{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\mathrm{TCC}}\mathcal{L}_{\mathrm{TCC}} + \lambda_{\mathrm{ICC}}\mathcal{L}_{\mathrm{ICC}}. \tag{31}$$

典型设置为：Δ=1；λ_{SD}=1.0；λ_v 热启（前 5k iter 从 0 线性升至 0.5）；$\rho^\star$ 从 1.0 升至 1.5、$\lambda_\rho\in[0.1, 0.2]$；$\lambda_{\mathrm{TCC}}\in[0.05, 0.1]$；$\lambda_{\mathrm{ICC}}\in[0.03, 0.07]$。在新视角/新时间渲染时，仅需前向使用学习到的 PVG 参数；若需可视化动静分离，可渲染点元级 w_i^{stat} 或速度门控 g_i 得到掩码，而无需教师特征。

F. 语义驱动的时间约束与一致性增强（Semantic Temporal Constraints & Consistency)

为降低静态背景的伪运动并维持动态目标的时间连贯性，本节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 所得的像素级静态概率 $\delta(\mathbf{p},t)$ 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），并在时序层约束几何与语义的一致性（TCC/ICC）。直观地，δ 回答“哪里更应静”，w_i^{stat} 回答“哪些点更应稳”，二者共同决定约束的施加强度。

a) 语义速度约束（SVC）： 道路、建筑与天空等背景应几乎静止，而车辆与行人允许产生运动。为此，以点元语义向量 $f_{\mathrm{sem},i}$ 产生速度门控并直接作用于 PVG 的速度基向量：

$$g_i = \sigma(\mathbf{w}_g^\top f_{\mathrm{sem},i} + b_g) \in (0,1), \qquad v_i^{\mathrm{eff}} = g_i\, v_i, \tag{32}$$

随后以 v_i^{eff} 替代 v_i 更新点元轨迹 $\mu_i(t)$（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：

$$V(\mathbf{p},t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,\big\|\Pi\big(\mu_i(t+\Delta)\big) - \Pi\big(\mu_i(t-\Delta)\big)\big\|_1. \tag{33}$$

训练时将 $\delta(\mathbf{p},t)$ 作为软静态权重抑制静区速度：

$$\mathcal{L}_v = \mathrm{mean}_{\mathbf{p}\in\Omega}\big(\delta(\mathbf{p},t)\cdot V(\mathbf{p},t)\big). \tag{34}$$

直观上，δ 越大（越像静态），对该像素的“制动”越强；前景目标因 δ 较小且门控 g_i 较大，从而保留合理运动。

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

$$\rho_i = \beta_i / l_i, \tag{35}$$

并对静态倾向更高的点元施加下界约束。令 w_i^{stat} 为由 δ 按可见性回投得到的点元静态概率，则

$$\mathcal{L}_\rho = \sum_i w_i^{\mathrm{stat}} \max\big(0,\ \rho^\star - \rho_i\big). \tag{36}$$

直观上，ρ_i 越大表示“活得更久、摆得更慢”，更接近静态；训练时将 $\rho^\star$ 由 1.0 线性升至 1.5 可逐步提升稳态。

c) 时间闭环一致性（TCC）： 为抑制时间维度上的累积偏移，要求点元在相邻时刻的往返映射闭环一致。记 $\Phi_i(\cdot; t_1 \to t_2)$ 为按当前时间参数将点元从 t_1 映射至 t_2 的算子，则

$$\mathcal{L}_{\mathrm{TCC}} = \frac{1}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} \Big\| \Phi_i^{-1}\big(\Phi_i(\mu_i;\, t{-}\Delta \to t{+}\Delta);\, t{+}\Delta \to t{-}\Delta\big) - \mu_i \Big\|_1, \tag{37}$$

其中集合 $\mathcal{I}$ 由可见性阈值筛选出遮挡轻微的点元。若参数正确，“前进再后退”应回到原位；否则该项将拉回偏移，抑制时间漂移（在遮挡边界可用 Huber 替代 L_1 提升鲁棒性）。

d) 实例一致性增强（ICC）： 动态对象在跨时间应保持语义稳定。为此，按实例维护语义原型 p_k（由其成员点元语义向量的指数滑动平均得到），并以中心式一致性损失使成员靠近原型：

$$\mathcal{L}_{\mathrm{ICC}} = \frac{1}{|\mathcal{S}_k|} \sum_{i\in\mathcal{S}_k} \|f_{\mathrm{sem},i} - p_k\|_2^2, \qquad p_k \leftarrow \eta\, p_k + (1-\eta)\, \frac{1}{|\mathcal{S}_k|} \sum_{i\in\mathcal{S}_k} f_{\mathrm{sem},i}, \tag{38}$$

其中 $\eta \in [0.7, 0.95]$ 控制更新速率，$\mathcal{S}_k$ 由实例分割与轻量跨帧关联获得。该项可避免“同一物体语义忽明忽暗”，从而使 SVC 的速度门控在实例内部更稳定。

e) 训练与整合： 训练期间，δ 始终以软权重形式参与式 (101)，并通过回投形成 w_i^{stat} 进入式 (103)；$\mathcal{L}_{\mathrm{TCC}}$ 与 $\mathcal{L}_{\mathrm{ICC}}$ 仅在可见性良好的样本上计算。总目标函数写作

$$\mathcal{L} = \mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{SD}}\mathcal{L}_{\mathrm{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\mathrm{TCC}}\mathcal{L}_{\mathrm{TCC}} + \lambda_{\mathrm{ICC}}\mathcal{L}_{\mathrm{ICC}}, \tag{39}$$

其中典型设置为：Δ=1；λ_{SD}=1.0；λ_v 在前 5k 次迭代从 0 线性升至 0.5；$\rho^\star$ 从 1.0 升至 1.5 且 $\lambda_\rho \in [0.1, 0.2]$；$\lambda_{\mathrm{TCC}} \in [0.05, 0.1]$；$\lambda_{\mathrm{ICC}} \in [0.03, 0.07]$。语义门控（式 (99)）仅引入一次点乘与 Sigmoid，额外开销可忽略；其余计算完全复用 3DGS/4DGS 的光栅化与可见性管线，因而实现简单、收敛稳定。

G. 语义驱动的时间约束（*Semantic Temporal Constraints*）

a) 问题动机与直觉： 仅依赖重建误差训练动态高斯时，街景视频中常出现两类典型失效：其一，**静态背景抖动/漂移**——相机在动但路、楼、天应近乎静止，却被误解释为物体运动；其二，**长期不稳**——即便瞬时外观重建良好，长序列中静态区域的点元仍可能缓慢游走。我们观察到：通过§3.3 得到的像素静态概率 $\delta(\mathbf{p}, t)$ 在路面/建筑等区域显著偏高，说明这些位置**应当接近零速度**；同时，能稳定支撑这些区域的点元，其时间参数也应体现**“活得更久、摆得更慢”**的性质。基于此，本文仅用两条轻量、可解释的约束即能显著改善时序稳定性：**语义速度约束（*SVC*）**与**静态寿命先验（*SLP*）**。

b) 语义速度约束（SVC）：静区近零速度： 对每个高斯点元 i，由其语义向量 $f_{\mathrm{sem},i}$ 产生一个速度门控，直接作用于 PVG 的速度基向量 v_i：

$$g_i = \sigma(\mathbf{w}_g^\top f_{\mathrm{sem},i} + b_g) \in (0,1), \qquad v_i^{\mathrm{eff}} = g_i\, v_i, \tag{40}$$

并以 v_i^{eff} 替代 v_i 更新轨迹 $\mu_i(t)$（其余渲染流程不变）。直觉上，**像“路/楼/天”的点元**使 $g_i \downarrow$、速度被抑制；**像“车/人”的点元**使 $g_i \uparrow$、保留合理运动。为度量瞬时残余运动，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：

$$V(\mathbf{p}, t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p}, t) \left\| \Pi\big(\mu_i(t+\Delta)\big) - \Pi\big(\mu_i(t-\Delta)\big) \right\|_1. \tag{41}$$

训练时不对 δ 二值化，而是直接将其作为“静态程度”的软权重抑制静区速度：

$$\mathcal{L}_v = \mathrm{mean}_{\mathbf{p}\in\Omega}\big(\delta(\mathbf{p}, t) \cdot V(\mathbf{p}, t)\big). \tag{42}$$

这样做的直觉很简单：δ 越大越该静，则对该像素的“制动”越强；动态目标因 δ 通常较小，门控 g_i 较大，从而允许其真实运动被保留。

c) 静态寿命先验（SLP）：长期稳态： 仅靠瞬时速度抑制仍可能导致长序列中的缓慢漂移。我们利用 PVG 的时间参数定义静态度

$$\rho_i = \beta_i / l_i, \tag{43}$$

其中 β_i 为寿命尺度、l_i 为周期长度；ρ_i 越大表示“活得更久/摆得更慢”，越接近静态。将像素静态概率 $\delta(\mathbf{p}, t)$ 按可

见性回投为点元级权重 w_i^{stat}（与§3.3 相同的加权聚合），对更应静的点元施加下界约束：

$$\mathcal{L}_\rho = \sum_i w_i^{\mathrm{stat}} \max\big(0,\ \rho^\star - \rho_i\big). \tag{44}$$

实践中将阈值 $\rho^\star$ 从 1.0 线性升至 1.5 可逐步增强稳态。直观地，SVC 约束瞬时速度，SLP 约束时间形态，二者分别对应“**当下别抖**”与“**长期别漂**”，效果互补。

d) 训练与实现: 训练期间，门控层参数 $(\mathbf{w}_g, b_g)$ 与主网络端到端优化（Adam），初始化为零以避免早期过抑制；像素静态概率 δ 来自§3.3 的可学习融合器，作为软权重参与式 (101) 并通过回投形成 w_i^{stat} 进入式 (103)。典型设置为 Δ=1 帧、λ_v 在前 5k 次迭代从 0 线性升至 0.5、$\lambda_\rho \in [0.1, 0.2]$。两项约束完全复用 3DGS/4DGS 的光栅化与可见性计算，仅在参数层与像素层增加常数级开销，因而实现简单、收敛稳定，并能显著减少背景伪动、提升长序列稳定性。

H. 优化（*Optimization*）

本节给出训练与工程实现的关键细节，目标是在不增加体渲复杂度的前提下，稳健、高效地联合优化重建、2D→4D 语义蒸馏与时间约束（SVC/SLP），并降低内存与 IO 开销。

a) 损失组成与调度: 总目标函数为

$$\mathcal{L} = \mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{SD}}\mathcal{L}_{\mathrm{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho, \tag{45}$$

其中 $\mathcal{L}_{\mathrm{rgb}}$ 采用 Charbonnier（或 $L_1 + \mathrm{SSIM}$）以增强对光照变化的鲁棒性；$\mathcal{L}_{\mathrm{SD}}$ 为§3.2 的像素级蒸馏；$\mathcal{L}_v$ 与 $\mathcal{L}_\rho$ 分别对应§3.4 的 SVC 与 SLP。为避免早期几何尚未成形时过强时间约束，我们对系数采用热启：λ_v 在前 5k iter 从 0 线性升至 0.5，$\lambda_\rho \in [0.1, 0.2]$ 固定或缓升；λ_{SD}=1.0。实证显示该调度能显著降低背景“早抑制”导致的过暗或细节缺失。

b) 分组参数化与学习率: 为兼顾稳定与收敛速度，按功能分组设不同比例学习率：几何（μ, Σ）与外观（SH/颜色）用基准 LR，时间参数（v, l, β, τ）用 0.5×，语义相关（$f_{\mathrm{sem}}, U(\cdot)$）与门控/融合器参数（$\mathbf{w}_g, b_g, a, b, c$）用 1.0×。优化器采用 Adam（$\beta_1$=0.9, β_2=0.99），权重衰减 10^{-6}；梯度裁剪到 $\ell_2 \le 1$ 以避免偶发极端视角导致的爆梯。

c) 可见性一致的像素采样: 每次迭代在瓦片优先的基础上混合采样：70% 随机瓦片、30% 来自“高差异/边界”区域（D 与图像梯度的并集）。该策略让蒸馏与重建在同一遮挡排序上“看”到难点像素，促使 f_{sem} 与几何/遮挡同步收敛，减少动态边界的拖影。

d) SVC/SLP 的数值稳健性: 速度门控初始化为“中性”以防早期过抑制：$\mathbf{w}_g$=$\mathbf{0}, b_g$=0 $\Rightarrow$ $g_i \approx 0.5$，随后由语义自适应学习；像素静态概率 $\delta(\mathbf{p}, t)$ 始终以软权重参与式 (101)，不二值化，并通过可见性回投生成点元级权重 w_i^{stat} 用于式 (103)。为防数值发散，l_i, β_i 用对数参数化（l_i=$\exp(\hat{l}_i)$，β_i=$\exp(\hat{\beta}_i)$），协方差以旋转四元数 q_i 与对角尺度 s_i 表达，确保 $\Sigma_i \succ 0$。

e) 教师特征的缓存与压缩: LSeg 教师特征 F_t 采用离线提取 + 轻量缓存：对原始分辨率进行 2∼4 倍下采样并存 FP16；如需进一步降 IO，可对通道做 PCA 压至 64/128 维，训练时上采样并经升维头 $U(\cdot)$ 对齐到教师维度计算蒸馏与差异 D。该流程在几乎不影响语义判别力的前提下，将特征存储压缩至原始的 1/4∼1/8。

f) 渲染与内存优化: 复用 3DGS 的瓦片化可微光栅化，在每瓦片仅保留*Top-K* 可见高斯（如 K=48），其余直接裁剪；同时对极小不透明度的点（o_i<10^{-3}）行程式剔除。语义分支与 RGB 共用排序，避免重复排序开销；语义升维 $U(\cdot)$ 以 1×1 卷积实现，计算量可忽略。训练端到端混合精度（AMP）与逐瓦片反向传播，最大化显存利用率。

g) 正则与防退化: 为防止语义向量过大或坍缩，引入轻量 ℓ_2 正则 $\sum_i \|f_{\mathrm{sem},i}\|_2^2$；对时间参数加入平滑项 $\|\Delta v_i\|_2^2$（跨若干迭代的 EMA 差分）以抑制抖动；可选地对像素速度图 V 加一个小幅 TV 正则（10^{-5}∼10^{-4}）消除孤立噪点。实践中，这些项对收敛稳定性帮助明显，而对质量影响可忽略。

h) 课程学习与多尺度: 采用“分辨率从低到高、曝光/颜色轻扰动”的课程：前 2k iter 用 1/2 尺度训练以快速成形几何与语义，随后切换到全分辨率；RGB 端做轻微亮度/对比度抖动，但教师特征固定从未增强的原图提取，避免教师端漂移。多尺度下，蒸馏在低尺度、重建在高尺度共同进行，可显著降低早期伪影。

i) 实现与复杂度: 全流程新增计算仅为：一层 1×1 升维、一次向量点乘与 Sigmoid（门控），以及基于已有投影的像素速度图统计；时间复杂度与显存开销均与 3DGS/4DGS 同量级。典型设置下（1024×576、单卡 24GB），每迭代处理 1∼2 帧、每帧 $N_{\mathrm{pix}} \approx 4$∼6 万像素，训练稳定、吞吐不降。

j) 小结: 上述优化从损失调度、分组学习率、可见性一致采样、教师特征压缩、瓦片化裁剪与数值正则等层面，配合 SVC/SLP 的软约束形态，使模型在几乎零额外复杂度下显著减少背景伪动与长序列漂移，并在城市街景等低视角多遮挡场景中获得稳定收敛与可复现实验结

果。

I. 优化（*Optimization)*

本节给出训练与工程实现的关键细节，采用阶段化的联合优化：先在静区主导的约束下稳定几何与语义，再启用时间项完善动态表现。整个流程不引入体渲复杂度的增加，所有附加项均复用 3DGS/4DGS 的光栅化与可见性计算。

a) 初始化与热启: 我们首先以静态 3DGS 进行短暂预热：用 RGB 重建损失在若干迭代内优化位置 μ、协方差 Σ、不透明度 o 与外观参数（SH），得到稳定的初始几何；随后扩展为 PVG，将速度 v 初始化为 0、周期 l 与寿命尺度 β 初始化为较大值，使模型从“近静止”出发；同时启动§3.2 的 2D→4D 语义蒸馏以写入点元语义向量 f_{sem}。像素静态概率 $\delta(\mathbf{p},t)$ 由§3.3 的融合器在线产生，但其参与时间项的权重在初期采用热启（见后文）。

b) 阶段 I：静区主导的几何/语义成形: 本阶段旨在在掩码与语义蒸馏的协同下，快速稳定静态背景的几何与外观，并学习到可靠的静态概率。对第 t 帧，渲染 $\hat{I}_t$，采用混合像素重建损失

$$\mathcal{L}_{\text{rgb}}(t) = (1-\lambda_{\text{ssim}})\,\|I_t - \hat{I}_t\|_1 + \lambda_{\text{ssim}}\,(1-\text{SSIM}(I_t,\hat{I}_t)), \tag{46}$$

并用静态软权重强调静区重建

$$\mathcal{L}^{\text{stat}}_{\text{rgb}} = \text{mean}_{\mathbf{p}\in\Omega}\left(\delta(\mathbf{p},t)\cdot\ell_{\text{pix}}(\mathbf{p},t)\right), \qquad \ell_{\text{pix}} \text{ 为式 (46) 的逐像素项。} \tag{47}$$

同时施加像素级语义蒸馏损失（§3.2）$\mathcal{L}_{\text{SD}}$，并加入瓦片/栅格化的 Total Variation（TV）正则以抑制高频噪声：

$$\mathcal{L}_{\text{stage I}} = \mathcal{L}^{\text{stat}}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}. \tag{48}$$

直觉上，式 (47) 让“更该静”的区域先收敛，蒸馏项将稳定的 2D 语义写入 4D 点元，从而为时间约束提供可靠的路由信号。

c) 阶段 II：时间约束与全量监督: 几何与语义成形后，启用语义速度约束（SVC）与静态寿命先验（SLP）（§3.4）。同时，为便于联合优化，我们用与 RGB 相同的 α-合成权重 $T_i\alpha_i$ 渲染深度、法线与像素速度图：

$$\{D,N,V\} = \sum_{i\in\mathcal{V}} T_i\,\alpha_i\,\{d_i,n_i,v_i\}, \qquad T_i = \prod_{j<i}(1-\alpha_j), \tag{49}$$

其中 d_i 为点到相机的深度投影，n_i 为法线（由高斯在像素处的主轴投影近似或由深度梯度估计），v_i 为§3.4 的像素位移项。若存在稀疏/投影深度监督（如 LiDAR），采用

$$\mathcal{L}_D = \|D - D^{\text{gt}}\|_1, \tag{50}$$

并可选加入法线一致性

$$\mathcal{L}_N = \text{mean}_{\mathbf{p}}(1 - \langle N(\mathbf{p}), N^{\text{gt}}(\mathbf{p})\rangle). \tag{51}$$

时间项采用§3.4 的

$$\mathcal{L}_v = \text{mean}_{\mathbf{p}}(\delta(\mathbf{p},t)\cdot V(\mathbf{p},t)), \qquad \mathcal{L}_\rho = \sum_i w_i^{\text{stat}}\max(0,\rho^\star-\rho_i), \tag{52}$$

其中 w_i^{stat} 由 δ 按可见性回投得到。阶段 II 的总损失为

$$\mathcal{L}_{\text{stage II}} = \mathcal{L}_{\text{rgb}}+\lambda_{\text{SD}}\mathcal{L}_{\text{SD}}+\lambda_v\mathcal{L}_v+\lambda_\rho\mathcal{L}_\rho+\lambda_D\mathcal{L}_D+\lambda_N\mathcal{L}_N+\lambda_{\text{tv}}\mathcal{L}_{\text{tv}}. \tag{53}$$

实践表明，SVC 抑制“当下抖动”，SLP 抑制“长序列慢漂”，两者与深度/法线等弱监督协同，能在不增加渲染复杂度的前提下显著提升时间稳定性。

d) 超参数与调度: 优化器采用 Adam（β_1=0.9, β_2=0.99），权重衰减 10^{-6}；按功能分组设学习率：几何/外观用基准 LR，时间参数（v,l,β,τ）用 0.5×，语义相关（$f_{\text{sem}}, U(\cdot)$）与门控/融合器参数（$\mathbf{w}_g, b_g, a, b, c$）用 1.0×。系数建议：$\lambda_{\text{ssim}}$=0.2，$\lambda_{\text{SD}}$=1.0，$\lambda_{\text{tv}}=10^{-5}$；阶段 II 中 λ_v 采用热启（前 5k iter 从 0 线性升至 0.5），$\lambda_\rho \in [0.1, 0.2]$，有深度监督时 $\lambda_D \in [0.5, 1.0]$，有法线监督时 $\lambda_N \in [0.1, 0.3]$。阈值 $\rho^\star$ 建议从 1.0 线性升至 1.5。

e) 工程细节与内存优化: 教师特征 F_t 采用离线 FP16 缓存，在 $H/2$ 或 $H/4$ 尺度存储；必要时对通道做 PCA 压至 64/128 维并在训练时经 $U(\cdot)$ 升维参与蒸馏与差异计算。渲染端复用瓦片化可微光栅化，并在每瓦片仅保留 Top-K 可见高斯（如 K=48）；对极小不透明度点（$o_i<10^{-3}$）行程式剔除。全流程采用 AMP 混合精度与逐瓦片反向传播以控制显存。

f) 小结: 本节给出的阶段化优化在“静区先稳—再加时间约束”的叙事下，将重建、蒸馏与语义驱动的时间正则紧密耦合：阶段 I 聚焦静区几何/语义成形，阶段 II 以 SVC/SLP 巩固时间稳定，并在可用时引入轻量几何监督（深度/法线）。该策略在城市街景的长序列与遮挡场景中表现稳健，且与 3DGS/4DGS 同量级的效率使其便于复现与扩展。

J. 优化（*Optimization*）

本节给出单阶段（Scheme A）+ 热启的完整训练与工程细节。核心原则是：重建与语义蒸馏始终开启，时间约束（*SVC/SLP*）按计划逐步升权；同时通过分组学习率、可见性一致采样与特征缓存，确保在不增加体渲复杂度的前提下实现稳定高效收敛。

a) 训练目标与调度（单阶段 + 热启）： 总目标函数为

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho, \tag{54}$$

其中 $\mathcal{L}_{\text{rgb}}$ 为重建项（Charbonnier 或 L_1+SSIM），$\mathcal{L}_{\text{SD}}$ 为§3.2 的像素级语义蒸馏，$\mathcal{L}_v$ 与 $\mathcal{L}_\rho$ 分别对应§3.4 的 SVC 与 SLP。从训练开始到结束，$\mathcal{L}_{\text{rgb}}$ 与 $\mathcal{L}_{\text{SD}}$ 始终开启；时间约束采用热启：

$$\lambda_v(t) = \begin{cases} \lambda_v^\star \cdot \dfrac{t}{T_v}, & 0 \le t < T_v, \\ \lambda_v^\star, & t \ge T_v, \end{cases} \qquad \lambda_\rho(t) = \lambda_\rho^\star, \qquad \rho^\star(t) = \rho_0 + (\rho_1 - \rho_0)\cdot \min\left(1, \frac{t}{T_\rho}\right), \tag{55}$$

推荐设置：λ_{SD}=1.0，$\lambda_v^\star$=0.5，T_v=5k iter；$\lambda_\rho^\star \in [0.1, 0.2]$ 固定；ρ_0=1.0 → ρ_1=1.5，T_ρ=30k iter。直觉上，先让几何与语义成形，再逐步让“静区近零速度/长期稳态”的约束接管时间行为，避免早期过抑制导致的发灰与细节损失。

b) 优化器与分组学习率： 采用 Adam（β_1=0.9, β_2=0.99，权重衰减 10^{-6}），梯度裁剪到 $\ell_2 \le 1$。为兼顾稳定与速度，按参数功能分组设学习率比例（以基准 LR 记为 η）：

- 几何与外观：$\{\mu, \Sigma, q, s, \text{SH}/c, o\}$ 用 η；
- 时间参数：$\{v, l, \beta, \tau\}$ 用 0.5η（更稳）；
- 语义与小头：$\{f_{\text{sem}}, U(\cdot), a, b, c, \mathbf{w}_g, b_g\}$ 用 η。

实践中 $\eta \in [1\times10^{-3}, 2\times10^{-3}]$ 表现稳定；速度门控初始化 $\mathbf{w}_g$=$\mathbf{0}, b_g$=0 使 $g_i \approx 0.5$（中性），融合器初始化 a=1, b=1, c=0。

c) 像素/瓦片采样与难例挖掘： 每迭代在可见瓦片上混合采样：70% 随机瓦片、30% 来自“高差异/高梯度”区域（由 D 与图像 Sobel 边缘的并集筛选）。该策略让蒸馏、重建与时间项在同一遮挡排序下关注难点像素（动态边界、遮挡区），促使语义与几何同步收敛。用于 $\mathcal{L}_v$ 的对称时间步取 Δ=1 即可。

d) 数值稳定与参数化： 时间尺度采用对数参数化以避免无效负值：l_i=$\exp(\hat{l}_i)$、β_i=$\exp(\hat{\beta}_i)$；协方差以旋转四元数 q_i 与对角尺度 s_i 表达，确保 $\Sigma_i \succ 0$。为防极端梯度导致门控饱和，训练中可对 g_i 做轻微夹紧（如 $g_i \in [0.05, 0.95]$，仅用于前向），并在像素速度图 V 的分母加 ε 以避免除零。重要的是：$\delta(\mathbf{p}, t)$ 始终以软权重参与 $\mathcal{L}_v$，不做二值化，并通过可见性回投形成 w_i^{stat} 用于 $\mathcal{L}_\rho$。

e) 教师特征的缓存与通道压缩： LSeg 特征 F_t 采用“离线提取 + 轻量缓存”：对原图下采样（$H/2$ 或 $H/4$）并以 FP16 存储；可选对通道做 PCA 压至 64/128 维以降 IO。训练时上采样并经 $U(\cdot)$ 对齐到教师维度参与蒸馏与差异 D 计算。该流程几乎不影响语义判别力，可将存储/带宽开销降至原始的 1/4∼1/8。

f) 渲染与内存优化： 复用 3DGS 的瓦片化可微光栅化，在每瓦片仅保留 Top-K 可见高斯（如 K=48），并对极小不透明度点（o_i<10^{-3}）做行程式剔除。语义分支与 RGB 共用排序，避免重复排序；语义升维 $U(\cdot)$ 用 1×1 卷积实现，计算量可忽略。训练采用混合精度（AMP）与逐瓦片反向，最大化显存利用率。

g) 课程学习与多尺度： 前 2k iter 以 1/2 分辨率快速成形几何与语义，随后切换全分辨率；RGB 端做轻微亮度/对比度扰动以增强鲁棒，但教师特征始终从未增强的原图提取，避免教师端漂移。多尺度下，蒸馏在低尺度、重建在高尺度共同进行，可显著降低早期伪影与抖动。

h) 训练循环概述： 一次迭代包含：渲染 $\hat{I}_t$ 与学生语义图 F_s（升维得 $\tilde{F}_s$）→ 计算蒸馏 $\mathcal{L}_{\text{SD}}$ 与重建 $\mathcal{L}_{\text{rgb}}$ → 由 $F_t, \tilde{F}_s$ 得差异 D，与 M_{sem} 融合得 δ → 计算像素速度图 V 并形成 $\mathcal{L}_v$ → 回投 δ 得 w_i^{stat} 并形成 $\mathcal{L}_\rho$ → 按式 (54) 与 (55) 组合总损失反传更新。

i) 复杂度与可复现设置： 上述优化仅在参数层与像素层增加常数级运算（1×1 升维、点乘+Sigmoid 门控与投影差分统计），总体时间/显存开销与 3DGS/4DGS 同量级。典型配置（1024×576、单卡 24GB）下，每迭代处理 1∼2 帧、每帧 N_{pix}≈4∼6 万像素可保持稳定吞吐。按推荐超参（λ_{SD}=1.0，$\lambda_v^\star$=0.5，$\lambda_\rho^\star$=0.15，T_v=5k，T_ρ=30k，η=1×10^{-3}）可直接复现实验结果；如出现早期背景发灰，适当减小 $\lambda_v^\star$ 或延长 T_v 即可。

K. 优化（*Optimization*）

a) 总体目标与权重调度： 训练目标由重建、语义蒸馏与时间约束三部分组成：

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{56}$$

其中 $\mathcal{L}_{\text{rgb}}$ 为照片重建损失（L_1+SSIM 的线性组合），$\mathcal{L}_{\text{SD}}$ 为§3.2 的特征蒸馏损失，$\mathcal{L}_v$ 与 $\mathcal{L}_\rho$ 分别为§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（见下）。为避免早期过度抑制动态与造成不稳定，我们采用温启动权重：λ_{SD}=1.0（常数），λ_v 在前 5k 次迭代从 0 线性升至 0.5，$\rho^\star$ 从 1.0 线性升至 1.5，$\lambda_\rho \in [0.1, 0.2]$。该调度使模型先收敛几何/外观，再逐步写入时间约束。

b) 训练日程与初始化： 我们采用单阶段端到端训练，但对时间相关参数做轻微延迟激活：初始化 v_i=$\mathbf{0}$、l_i=1、β_i=1，并在第 1k 次迭代后放宽其学习率（见下）以允许合理运动；速度门控层参数 $(\mathbf{w}_g, b_g)$ 与融合器 (a, b, c) 均初始化为 0（$g_i \approx 0.5$、$\delta \approx 0.5$ 的中性状态），避免早期

过抑制。语义向量 $f_{\text{sem},i}$ 以小幅高斯噪声初始化并在 $\mathcal{L}_{\text{reg}}$ 中施加 ℓ_2 正则，以防特征爆炸：

$$\mathcal{L}_{\text{reg}} = \lambda_f \sum_i \|f_{\text{sem},i}\|_2^2 + \lambda_v^{\ell_2} \sum_i \|v_i\|_2^2, \qquad \lambda_f{=}1\text{e}{-}4,\ \lambda_v^{\ell_2}{=}5\text{e}{-}5. \tag{57}$$

c) 像素与时间采样: 每次迭代从帧集合中采样中心时刻 t 及对称时间步 $t \pm \Delta$（Δ=1），并在图像上采用瓦片化随机像素采样（与 3DGS 相同）。为避免静区/动区极度不平衡，我们进行分层采样：按上一轮的静态概率 $\delta(\mathbf{p}, t)$，从“高δ”（静倾向）与“低δ”（动倾向）区域各采样一半像素，用于计算 $V(\mathbf{p}, t)$ 与蒸馏/重建损失；该做法能显著稳定 $\mathcal{L}_v$ 的梯度，减少训练后期的震荡。

d) 特征缓存与内存优化: 教师特征 F_t=LSeg(I_t) 采用离线缓存：以 FP16 在 $H/2$ 或 $H/4$ 尺度存储，并可选地做 PCA 压缩至 64/128 维；训练时上采样并经 1×1 对齐头 $U(\cdot)$ 投到教师维度参与蒸馏与差异计算。渲染与蒸馏均在随机瓦片上进行，配合混合精度（FP16/bfloat16）与梯度累积即可在单卡显存内完成较高分辨率训练。

e) 密度调整与稀疏化: 密度调整遵循 3DGS 的分裂/剪枝策略：当某点在屏幕空间的覆盖半径大且对重建贡献高时进行分裂；当其长期不透明度低、贡献小或处于重复冗余区域时剪枝。为降低“点元轮换”的风险，我们在静态概率高的区域（高 w_i^{stat}）降低分裂频率、提高剪枝阈值，鼓励由少量、长寿命点承载背景；在动态区域则保持默认分裂以充分表达细节。

f) 优化器与学习率: 除相机/外参固定外，其余参数均使用 Adam 优化。推荐初始学习率：中心 μ_i 与尺度/旋转 s_i, q_i 为 1e−3，不透明度/着色参数为 2e−3，语义向量 $f_{\text{sem},i}$ 与升维头 U 为 1e−3，时间参数 (v_i, l_i, β_i) 为 5e−4（在第 1k 次迭代升至 1e−3），门控与融合器参数为 5e−4。对所有参数采用余弦退火至初值的 1/10；全局梯度裁剪为 1.0 以避免偶发的爆梯度。

g) 鲁棒性细节: 在 $V(\mathbf{p}, t)$ 的计算中，遮挡边界最易产生尖峰梯度。我们在两处加入稳健化：其一，在像素速度的 L_1 中使用 Charbonnier 近似（$\sqrt{x^2+\epsilon^2}$，ϵ=1e−3）；其二，对可见性权重 $w_i(\mathbf{p}, t)$ 设置最小阈值屏蔽极小贡献的远端点。对 $\mathcal{L}_{\text{rgb}}$，加入轻度颜色抖动与亮度归一化以提升对曝光变化的鲁棒性，但不对输入做几何扰动以免破坏几何收敛。

h) 实现与开销: 上述优化完全复用 3DGS/4DGS 的光栅化与可见性计算；新增的计算仅为：一层线性门控（式 (99)）、一次对称时间差分（式 (100)）与按像素静态概率的加权（式 (101)、(103)）。实际测得，这部分带来的时间开销可忽略，而对背景伪动抑制与长序列稳定性的收益显著。

i) 为什么需要这一节（简要引导）: 训练目标需要同时兼顾外观/几何的正确重建、来自 2D 教师的语义约束，以及时间上的稳定性约束。经验上，若在几何尚未收敛时过早加强时间项，容易导致训练抖动或把真实运动“压没”。因此我们采用温启动：先让重建与语义蒸馏把几何与语义对齐，再逐步提高时间相关损失的权重。

j) 写法 A（推荐）：保留 $\mathcal{L}_{\text{rgb}}$ 记号，并给出其明确定义: 为保持总目标简洁清晰，我们在总式中使用 $\mathcal{L}_{\text{rgb}}$，并在此处一次性给出其组成（L_1 与 SSIM 的线性组合）：

$$\mathcal{L}_{\text{rgb}} = \lambda_1 \underbrace{\frac{1}{|\Omega|} \sum_{\mathbf{p}\in\Omega} \left\|\hat{I}_t(\mathbf{p}) - I_t(\mathbf{p})\right\|_1}_{\mathcal{L}_{L1}} + \lambda_{\text{ssim}} \underbrace{\Big(1 - \text{SSIM}(\hat{I}_t, I_t)\Big)}_{\mathcal{L}_{\text{SSIM}}}, \tag{58}$$

其中 $\hat{I}_t$ 是渲染图像，I_t 为真实图像，Ω 为可见像素集合。随后，总目标写为

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{59}$$

权重调度： λ_{SD}=1.0；λ_v 在前 5k 次迭代由 0 线性升至 0.5；$\rho^\star$ 从 1.0 线性升至 1.5 并配合 $\lambda_\rho \in [0.1, 0.2]$；$\lambda_1, \lambda_{\text{ssim}}$ 可取 (1.0, 0.2) 或据数据集微调。该设置使模型先收敛几何与语义，再逐步写入时间约束。

k) 写法 B（可选）：不使用 $\mathcal{L}_{\text{rgb}}$ 记号，直接展开到总式: 若希望在一处给出完全展开的目标，可直接写为

$$\mathcal{L} = \lambda_1 \frac{1}{|\Omega|} \sum_{\mathbf{p}\in\Omega} \left\|\hat{I}_t(\mathbf{p}) - I_t(\mathbf{p})\right\|_1 + \lambda_{\text{ssim}} \Big(1 - \text{SSIM}(\hat{I}_t, I_t)\Big) + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v \cdots \tag{60}$$

其余权重调度与上式相同。

建议： 写法 A 更利于阅读与引用（主文中保持总式紧凑，只需在本节给出 $\mathcal{L}_{\text{rgb}}$ 的一次性定义）；写法 B 适合在附录或消融对比中“逐项增减”时直接对照。

L. 优化（*Optimization*）

a) 设计动机与总体思路: 训练目标需要同时兼顾外观重建、来自 2D 教师的语义对齐，以及时序稳定性。经验上，若在几何尚未收敛时过早强化时间项，容易把真实运动“压没”或引入训练震荡。为此，我们采用温启动的权重调度：先让重建与语义蒸馏稳定几何与外观，再逐步加大时间相关约束（§3.4）在损失中的占比。这样可在不增加体渲复杂度的前提下，稳健写入“静区近零速度、长期不漂移”的先验。

b) 重建与总损失（最简形式）： 我们采用逐像素 L_1 光度项作为重建损失，并与语义蒸馏及时间约束共同优化：

$$\mathcal{L}_{\text{rgb}} = \frac{1}{|\Omega|} \sum_{\mathbf{p}\in\Omega} \left\| \hat{I}_t(\mathbf{p}) - I_t(\mathbf{p}) \right\|_1, \tag{61}$$

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{62}$$

其中 $\mathcal{L}_{\text{SD}}$ 为§3.2 的特征蒸馏，$\mathcal{L}_v, \mathcal{L}_\rho$ 为§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（见下）。

c) 权重与温启动: 默认设置：λ_{SD}=1.0；λ_v 在前 5,000 次迭代从 0 线性升至 0.5；阈值 $\rho^\star$ 从 1.0 线性升至 1.5，配合 $\lambda_\rho \in [0.1, 0.2]$。该调度使模型先“看清楚”（几何/外观），再“动得稳”（时间约束）。

d) 一次迭代的训练流程: 每一步更新按如下顺序进行（与 3DGS/4DGS 的光栅化管线一致）：

1) **采样与缓存检索：** 从训练序列采样中心时刻 t 及对称步 $t \pm \Delta$（默认 Δ=1），在图像上做瓦片化随机采样。加载（或在线计算）教师特征 $F_t = \text{LSeg}(I_t)$（推荐离线缓存至 FP16，$H/2$ 或 $H/4$ 尺度）。
2) **前向渲染（学生）：** 渲染 $\hat{I}_t$ 与学生语义图 F_s，经 1×1 线性头得 $\tilde{F}_s$（§3.2）。
3) **动静线索与融合：** 计算教师—学生特征差异 $D(\mathbf{p},t)$=1−cos$(\tilde{F}_s, F_t)$ 与语义先验 M_{sem}，经两输入逻辑回归得到像素静态概率

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

并按可见性回投形成点元静态权重 w_i^{stat}（§3.3）。
4) **时间约束：** 用语义门控 $g_i = \sigma(\mathbf{w}_g^\top f_{\text{sem},i} + b_g)$ 更新 v_i^{eff}=$g_i v_i$，计算像素速度图 $V(\mathbf{p},t)$ 并得到

$$\mathcal{L}_v = \text{mean}_{\mathbf{p}}\left(\delta(\mathbf{p},t)\cdot V(\mathbf{p},t)\right), \qquad \mathcal{L}_\rho = \sum_i w_i^{\text{stat}} \max(0, \rho^\star - \rho_i).$$

5) **重建与蒸馏：** 计算 $\mathcal{L}_{\text{rgb}}$ 与 $\mathcal{L}_{\text{SD}}$，并加上正则 $\mathcal{L}_{\text{reg}}$。
6) **密度自适应：** 按照 3DGS 的分裂/剪枝策略更新点集；在高 w_i^{stat} 区域降低分裂频率、提高剪枝阈值，鼓励“少量且长寿命”的静态承载。
7) **反传与更新：** 采用 Adam 对所有可学习参数（几何、外观、语义向量 f_{sem}、升维头 U、时间参数 v, l, β、融合器 a, b, c、门控 $\mathbf{w}_g, b_g$）联合更新。使用混合精度与梯度裁剪（1.0）提升稳定性。

e) 采样与稳定技巧: 为缓解动/静不平衡，我们对像素做*分层采样*：依据上一轮的 δ，从“高 δ（静倾向）”与“低 δ（动倾向）”区域各采一半，用于 $V(\mathbf{p},t)$、蒸馏与重建。为抑制遮挡处尖峰梯度，$V(\cdot)$ 的 L_1 可用 Charbonnier 近似 $\sqrt{x^2+\epsilon^2}$（ϵ=10^{-3}）替代；同时设置可见性权重 $w_i(\mathbf{p},t)$ 的最小门槛，忽略极小贡献的远端点。

f) 正则与初始化: 加入轻量正则

$$\mathcal{L}_{\text{reg}} = \lambda_f \sum_i \|f_{\text{sem},i}\|_2^2 + \lambda_v^{\ell_2} \sum_i \|v_i\|_2^2, \quad \lambda_f\text{=1e}-4,\ \lambda_v^{\ell_2}\text{=5e}-5. \tag{63}$$

初始化建议：v_i=**0**、l_i=1、β_i=1；门控与融合器权重全零（g_i≈0.5、δ≈0.5 的中性状态），在第 1,000 次迭代后提高时间参数学习率允许合理运动。

g) 效率与可复现性: 教师特征离线缓存（FP16 + 下采样 + 可选 PCA 到 64/128 维）显著降低 IO/显存；训练采用随机瓦片、混合精度与梯度累积可在单卡完成高分辨率训练。固定随机种子、记录权重调度与密度操作（分裂/剪枝）触发阈值，即可保证复现实验过程。

h) 重建项（简化写法）:

$$\mathcal{L}_{\text{rgb}} = \lambda_1\,\mathcal{L}_{L1} + \lambda_{\text{ssim}}\,\mathcal{L}_{\text{SSIM}}, \tag{64}$$

$$\mathcal{L}_{L1} = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}\left|\hat{I}_t(\mathbf{p}) - I_t(\mathbf{p})\right|, \qquad \mathcal{L}_{\text{SSIM}} = 1-\text{SSIM}(\hat{I}_t, I_t). \tag{65}$$

M. 优化（*Optimization)*

a) 总体思路: 训练目标需同时兼顾外观/几何重建、来自 2D 教师的语义对齐以及时间上的稳定性。经验上，若在几何尚未充分收敛时过早强化时间约束，易造成训练抖动或将真实运动过度抑制。因此我们采用*温启动*策略：先以重建与语义蒸馏稳固外观与语义，再逐步注入时间约束项的权重。

b) 重建项（简化写法）:

$$\mathcal{L}_{\text{rgb}} = \lambda_1\,\mathcal{L}_{L1} + \lambda_{\text{ssim}}\,\mathcal{L}_{\text{SSIM}}, \tag{66}$$

$$\mathcal{L}_{L1} = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}\left|\hat{I}_t(\mathbf{p}) - I_t(\mathbf{p})\right|, \qquad \mathcal{L}_{\text{SSIM}} = 1-\text{SSIM}(\hat{I}_t, I_t). \tag{67}$$

其中 $\hat{I}_t$ 为渲染图像，I_t 为真实图像，Ω 为当前可见像素集合。

c) 总目标与权重调度:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{68}$$

$\mathcal{L}_{\text{SD}}$ 为§3.2 的语义蒸馏损失，$\mathcal{L}_v$、$\mathcal{L}_\rho$ 分别为§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（见下）。权重采用温启动：λ_{SD}=1.0；λ_v 在前 5k 次迭代由 0 线性升至 0.5；阈值 $\rho^\star$ 从 1.0 线性升至 1.5，配合 $\lambda_\rho \in [0.1, 0.2]$；重建权重如 λ_1=1.0, λ_{ssim}=0.2（可随数据集微调）。

d) 正则化与初始化: 为防止语义向量与速度基过大，我们加入 ℓ_2 正则：

$$\mathcal{L}_{\text{reg}} = \lambda_f \sum_i \|f_{\text{sem},i}\|_2^2 + \lambda_v^{\ell_2} \sum_i \|v_i\|_2^2, \tag{69}$$

默认 $\lambda_f{=}1{\times}10^{-4}$，$\lambda_v^{\ell_2}{=}5{\times}10^{-5}$。时间相关参数初始化为中性：$v_i{=}\mathbf{0}$, $l_i{=}1$, $\beta_i{=}1$；语义门控与融合器参数初始化为零（$g_i{\approx}0.5$, $\delta{\approx}0.5$），避免早期过抑制。

e) 采样与效率: 每次迭代在时刻 t 及对称步 $t{\pm}\Delta$（$\Delta{=}1$）进行瓦片化像素采样；为缓解动静不平衡，按上一轮静态概率 δ 进行分层采样（静/动各占一半）。教师特征 F_t 采用离线缓存（FP16，$H/2$ 或 $H/4$，可选 PCA 至 64/128 维），训练时上采样并经 $1{\times}1$ 对齐头参与蒸馏与差异计算。全流程复用 3DGS/4DGS 的可微光栅化与可见性管线，新增计算主要为线性门控与对称时间差分，额外开销可忽略。

f) 优化器与学习率: 除固定外参外，其余参数均使用 Adam 优化，典型初始学习率：几何（位置/尺度/旋转）$1{\times}10^{-3}$，外观与对齐头 $1{\times}10^{-3}{\sim}2{\times}10^{-3}$，时间参数与门控/融合器 $5{\times}10^{-4}$（第 1k 次迭代升至 $1{\times}10^{-3}$）。采用余弦退火至初值的 $1/10$，全局梯度裁剪阈值 1.0。上述设置在不增加体渲复杂度的前提下，稳定地提升时序一致性并减小背景伪动。

II. 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），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。每个基元携带位置、各向异性尺度与旋转、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

$$G_i(\boldsymbol{x}) = \exp\Big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\Big). \tag{70}$$

其中，$\boldsymbol{\mu}_i$ 表示基元在三维空间中的位置；$\boldsymbol{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^\top\mathbf{R}_i^\top$ 为协方差矩阵，描述形状与方向，通常由旋转矩阵 $\mathbf{R}_i$ 与尺度矩阵 $\mathbf{S}_i$ 参数化。将三维高斯投影到像平面得到 2D 高斯，其投影协方差近似为

$$\boldsymbol{\Sigma}_i' = \mathbf{J}\,\mathbf{W}\,\boldsymbol{\Sigma}_i\,\mathbf{W}^\top\mathbf{J}^\top, \tag{71}$$

其中 $\mathbf{W}$ 是世界到相机的外参变换，$\mathbf{J}$ 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：

$$C = \sum_{i=1}^{N} T_i\,\alpha_i\,c_i, \qquad T_i = \prod_{j<i}(1-\alpha_j), \tag{72}$$

其中 α_i 由点元不透明度与其投影 2D 协方差在该像素的覆盖贡献共同决定，c_i 为外观。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 在表述上默认静态：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/幅度向量 $\boldsymbol{v}$、寿命尺度 β，定义

$$\tilde{\boldsymbol{\mu}}(t) = \boldsymbol{\mu} + \frac{l}{2\pi}\sin\Big(2\pi\frac{t-\tau}{l}\Big)\,\boldsymbol{v}, \qquad \tilde{o}(t) = o\cdot\exp\Big(-\tfrac{1}{2}(t-\tau)^2\,\beta^{-2}\Big). \tag{73}$$

时刻 t 的点元状态为 $H(t) = \{\tilde{\boldsymbol{\mu}}(t), \mathbf{q}, \mathbf{s}, \tilde{o}(t), \mathbf{c}\}$，整幅图像按

$$\hat{I}_t = \text{Render}\big(\{H_i(t)\}_{i=1}^{N};\, \mathbf{K}_t,\, \mathbf{E}_t\big) \tag{74}$$

渲染，其中 $\mathbf{K}_t, \mathbf{E}_t$ 分别为相机内参与外参。为衡量点元“静态度”，定义

$$\rho = \beta/l, \tag{75}$$

ρ 越大表示寿命相对周期更长、越趋近静止；当 $\boldsymbol{v}=\mathbf{0}$ 且 $\rho\to\infty$ 时，PVG 退化回标准 3DGS。以上设计使静态/动态以统一形式出现，仅通过 $\{\boldsymbol{v}, \beta, l, \tau\}$ 的取值区分。综上，PVG 以最小改动继承了 3DGS 的高效与可扩展性，又补足了动态建模与可编辑性，是面向道路环境的更合适表征选择。

B. *2D→4D* 语义蒸馏

仅凭重建损失难以稳定区分“相机运动”与“真实世界运动”。为增强模型对场景语义的理解，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。

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

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

学生端为每个高斯基元赋予可学习语义向量 $f_{\text{sem},i}$。与 RGB 渲染一致，像素 $\mathbf{p}$ 处的学生语义由可见性权重加权聚合；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

$$F_s(\mathbf{p},t)=\sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\, f_{\text{sem},i}. \tag{76}$$

为与教师通道数对齐，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：

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

采用像素级 L_1 蒸馏使学生贴近教师，构建蒸馏损失

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

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}，形成可随 PVG 形变在时间上搬运的 4D 语义表征。

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 $\tilde{F}_s$。本节据此构造由两类证据共同约束的帧级动静掩码：教师—学生特征差异与语义先验。前者反映像素处的语义不一致，后者提供先验上应静止的区域指示，二者经轻量融合得到稳定的静态掩码 M_{stat} 与动态掩码 M_{dyn}。

a) (a) 教师—学生特征差异: 当像素对应静态背景时，多帧观察指向同一世界点，学生的语义渲染应与教师一致；当像素位于动态对象或遮挡边界时，更易产生偏差。定义像素级余弦不相似度：

$$D(\mathbf{p},t)=1-\cos\big(\tilde{F}_s(\mathbf{p},t),\,F_t(\mathbf{p})\big). \tag{77}$$

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) 语义先验与融合: 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：

$$M_{\text{sem}}(\mathbf{p},t)=\sum_{k\in\mathcal{C}_{\text{stat}}} S_k(\mathbf{p},t). \tag{78}$$

为避免手工权重，用两输入一输出的逻辑回归融合两条证据并输出静态概率：

$$\delta(\mathbf{p},t)=\sigma\big(a\cdot(1-D(\mathbf{p},t))+b\cdot M_{\text{sem}}(\mathbf{p},t)+c\big), \tag{79}$$

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，δ 作为软静态权重直接用于§3.4 的时间约束；当需要导出或评测掩码时，再以固定阈值二值化：

$$M_{\text{stat}}(\mathbf{p},t)=\mathbf{1}\big(\delta(\mathbf{p},t)>\tau_s\big),\qquad \tau_s=0.5, \tag{80}$$

并令 $M_{\text{dyn}}(\mathbf{p},t)=1-M_{\text{stat}}(\mathbf{p},t)$。为降低误检，我们在实现中采用保守融合策略（静态取交、动态取并），此处不再展开额外公式。

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：

$$g_i=\sigma(\mathbf{w}_g^{\top} f_{\text{sem},i}+b_g)\in(0,1),\qquad \boldsymbol{v}_i^{\text{eff}}=g_i\,\boldsymbol{v}_i, \tag{81}$$

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：

$$V(\mathbf{p},t)=\sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,\big\|\Pi\big(\boldsymbol{\mu}_i(t+\Delta)\big)-\Pi\big(\boldsymbol{\mu}_i(t-\Delta)\big)\big\|_1. \tag{82}$$

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：

$$\mathcal{L}_v=\text{mean}_{\mathbf{p}\in\Omega}\big(\delta(\mathbf{p},t)\cdot V(\mathbf{p},t)\big). \tag{83}$$

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

$$\rho_i=\beta_i/l_i, \tag{84}$$

并用 w_i^{stat} 对静态倾向更高的点元施加下界约束：

$$\mathcal{L}_\rho=\sum_i w_i^{\text{stat}}\,\max\big(0,\ \rho^\star-\rho_i\big), \tag{85}$$

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

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda_{\text{SD}}\mathcal{L}_{\text{SD}} + \lambda_v\mathcal{L}_v + \lambda_\rho\mathcal{L}_\rho + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{86}$$

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：

$$\mathcal{L}_{\text{reg}} = \lambda_f \sum_i \|f_{\text{sem},i}\|_2^2 + \lambda_v^{\ell_2} \sum_i \|\boldsymbol{v}_i\|_2^2, \tag{87}$$

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。

III. 实验（EXPERIMENTS)

本章评估所提 SiT-PVG 在城市街景动态重建上的表现，并给出消融、效率与鲁棒性分析。我们关注三个核心问题：*(i)* 重建质量是否可比或优于现有方法？*(ii)* 静态背景的伪运动能否显著抑制，同时保留前景真实运动？*(iii)* 所引入的语义驱动时间约束在开销与稳定性上的收益如何？

A. 数据集与评测协议

数据集。 我们在 *KITTI-360* 与 *Waymo Open* 上进行评测。每条序列提供多视角或单目视频与标定参数；我们遵循既有划分，选取街景中含有丰富动态目标（车/人）与大尺度静态背景（路/楼/天）的片段进行训练与测试。

训练/测试划分。 每序列按时间切分为训练段与评测段，避免帧重叠；评测段既包含训练视角的已见时刻也包含未见时刻/新视角，以衡量插值与外推能力。除非特别说明，训练均在原始时间采样率下进行。

相机与预处理。 使用官方相机内外参，图像按长边缩放至 $\{640, 800\}$ 并进行轻度颜色归一化；不做几何抖动以保证几何可学习性。

B. 实现细节

教师特征缓存。 LSeg 在训练帧上离线前向得到 F_t，以 FP16 在 $H/2$ 尺度缓存；可选进行 PCA 降到 128 维，训练时上采样并用 1×1 线性头 $U(\cdot)$ 对齐至教师维度参与蒸馏与差异计算。

优化配置。 采用 Adam，全局梯度裁剪 1.0，余弦退火至初值 1/10。温启动调度：λ_v 在前 5k 迭代由 $0 \to 0.5$ 线性上升；$\rho^\star$ 在前 15k 由 $1.0 \to 1.5$ 线性上升；$\lambda_\rho \equiv 0.15$（或前 2k 预热到 0.15）。时间相关参数中性初始化（$\boldsymbol{v} = \mathbf{0},\ l = 1,\ \beta = 1$），语义门控/融合器线性层零初始化（$g \approx 0.5,\ \delta \approx 0.5$）。

采样与批处理。 每迭代从时刻 t 及对称步 $t\pm\Delta$（Δ = 1）采样瓦片；为缓解动静不平衡，按上一轮 $\delta(\mathbf{p}, t)$ 做分层像素采样（静/动各半）用于计算 $V(\mathbf{p}, t)$ 与重建/蒸馏项。

C. 比较方法（*Baselines)*

我们选取代表性的静态/动态高斯与语义增强方法作为对比：**3DGS**（静态高效基线）、**4DGS**（HexPlane 时间编码）、**PVG**（仅振荡参数化，无语义约束）、**Feature 3DGS**（语义特征蒸馏，静态）、**CoDa-4DGS**（语义增强的 4D 表征）。为公平起见，所有方法使用相同的相机与训练帧，必要时统一分辨率与训练步数；对需要语义的基线，使用相同的 LSeg 教师。

D. 评测指标（*Metrics)*

重建质量。 *PSNR/SSIM/LPIPS*（全图），并报告动态区域 *PSNR*（D-PSNR）与静态区域 *PSNR*（S-PSNR），区域由掩码 $M_{\text{dyn}}/M_{\text{stat}}$ 提供。

静区伪运动。 以§3.4 的像素速度图 $V(\mathbf{p}, t)$ 定义静态速度残差（SVR）：

$$\text{SVR} = \frac{1}{|\Omega_{\text{stat}}|} \sum_{\mathbf{p}\in\Omega_{\text{stat}}} V(\mathbf{p}, t), \quad \Omega_{\text{stat}} = \{\mathbf{p} \mid \delta(\mathbf{p}, t) \geq \tau_s\}.$$

数值越小越好，刻画背景抖动是否被抑制。

时间稳定性。 我们采用两种无 GT 指标：（i）像素级时域抖动：相邻时刻渲染的差分 $\|\hat{I}_{t+\Delta} - \hat{I}_{t-\Delta}\|_1$ 在静态区域的均值；（ii）轨迹闭环误差（TCE）：用 PVG 参数做一次 $t \to t{+}\Delta \to t$ 的往返映射平均偏差（与§3.4 的公式同形，作为度量而非损失）。两者越小越好。

效率与规模。 记录每序列的点元数量、渲染 FPS、显存占用、教师缓存大小与 I/O 吞吐，分析与基线的开销对比。

E. 主结果（Quantitative Results)

表 IV 与表 III 汇报在 KITTI-360 与 Waymo 上的主结果。SiT-PVG 在全图重建（PSNR/SSIM/LPIPS）上与强基线保持同量级，同时在 S-PSNR 与 SVR 上显著优于仅靠重建或仅靠时间编码的方法，表明静区伪运动得到有效抑制；在 D-PSNR 上保持与 4D 基线相当，说明前景真实运动未被过度抑制。

F. 实验设置（Experimental Setup)

本节给出我们的数据集、评测指标、对比基线与实现细节，确保实验可复现与公平对比。

a) 数据集: 我们在 **KITTI-360** 与 **Waymo Open** 的城市街景片段上评测。均使用官方提供的相机内参与外参，剔除失配与严重运动模糊帧。为覆盖“静态大场景 + 丰富动态体”的典型路况，每条序列选取包含车辆/行人/骑行者与大尺度背景（路/楼/天）的连续时间片段作为训练集与测试集。图像按长边缩放至 $\{640, 800\}$（默认 800），做亮度/对比度轻度归一化，不施加几何增强以保证几何可学习性。训练/测试采用时间不重叠的连续片段划分；测试同时包含训练视角的已见时刻与未见时刻/新视角两类，以分别评估插值与外推能力。除非另行说明，时间步取 Δ=1。

b) 评测指标: 我们报告三类指标：重建质量、静区伪运动与时间稳定性。 (i) **重建质量：**全图 PSNR/SSIM/LPIPS，并基于动静掩码给出静态区域 *PSNR*（S-PSNR）与动态区域 *PSNR*（D-PSNR）。 (ii) **静区伪运动：**使用§3.4 的像素速度图 $V(\mathbf{p}, t)$ 计算静态速度残差（SVR）：

$$\mathrm{SVR} = \frac{1}{|\Omega_{\mathrm{stat}}|} \sum_{\mathbf{p}\in\Omega_{\mathrm{stat}}} V(\mathbf{p}, t), \quad \Omega_{\mathrm{stat}} = \{\mathbf{p} \mid \delta(\mathbf{p}, t) \geq \tau_s\},$$

数值越小说明背景越稳定。 (iii) **时间稳定性：**报告像素级时域抖动 $\frac{1}{|\Omega_{\mathrm{stat}}|}\sum_{\mathbf{p}\in\Omega_{\mathrm{stat}}} \|\hat{I}_{t+\Delta}(\mathbf{p}) - \hat{I}_{t-\Delta}(\mathbf{p})\|_1$，以及轨迹闭环误差（TCE），后者以 PVG 参数做 $t \to t{+}\Delta \to t$ 的往返映射平均偏差作为度量（与§3.4 同形，用作指标而非损失）。所有指标均在同一分辨率与相机标定下计算。

c) 对比基线: 为公平评估，我们选取代表性方法并统一相机、训练帧与训练步数：**3DGS**（静态高效基线，显式点元 + 可微光栅化）； **4DGS**（在 3DGS 上引入时间编码/轻量 MLP 的 4D 表征）； **PVG (w/o SD)**（仅含周期振荡的时间参数化，不含语义蒸馏与时间约束）； **Feature 3DGS**（静态语义蒸馏到 3DGS）； **CoDa-4DGS**（含 2D 语义监督的 4D 表征）。对依赖语义的基线，均使用相同的 **LSeg** 作为教师模型，并在相同分辨率下缓存教师特征，以剥离语义质量差异的影响。

d) 实现细节: 所有实验基于 PyTorch 实现，默认在 **1× NVIDIA RTX 4090 (24GB)** 运行；长序列复现实验可在 **1× A100 (80GB)** 上等效运行。混合精度采用 FP16/BFloat16。每次迭代从时刻 t 及对称步 $t \pm \Delta$ 采样瓦片像素，按上一轮静态概率 $\delta(\mathbf{p}, t)$ 做分层采样（静/动各半）以稳定 $\mathcal{L}_v$ 的梯度。

*教师特征缓存与通道对齐：*对训练帧离线前向 LSeg 得到 F_t，以 FP16 在 $H/2$ 或 $H/4$ 尺度缓存，并可选做 PCA 压缩至 $\{64, 128\}$ 维；训练时上采样并经 1×1 线性头 $U(\cdot)$ 对齐至教师维度参与蒸馏与差异计算。

*参数初始化与正则：*时间相关参数采用中性初始化 $\boldsymbol{v}_i$=$\mathbf{0}$, l_i=1, β_i=1；语义门控与动静融合器的线性层参数置零，使 $g_i \approx 0.5$, $\delta \approx 0.5$（中性输出），避免早期过抑制。语义向量 $f_{\mathrm{sem},i}$ 以 $\mathcal{N}(0, 10^{-2})$ 初始化，并在 $\mathcal{L}_{\mathrm{reg}}$ 中施加 ℓ_2 正则（λ_f=1×10^{-4}）；速度基向量加 ℓ_2 正则（$\lambda_v^{\ell_2}$=5×10^{-5}）。

*优化器与调度：*除相机外参固定外，其余参数使用 Adam（β_1=0.9, β_2=0.999）。初始学习率：几何（位置/尺度/旋转）1×10^{-3}，外观与对齐头 $1 \sim 2 \times 10^{-3}$，时间参数与门控/融合器 5×10^{-4}（第 1k 次迭代升至 1×10^{-3}）。全局梯度裁剪阈值 1.0，余弦退火至初值的 1/10。权重温启动：$\lambda_v(n)$=$0.5 \cdot \min(\frac{n}{5000}, 1)$，$\rho^\star(n)$=$1.0 + 0.5 \cdot \min(\frac{n}{15000}, 1)$，$\lambda_\rho \equiv 0.15$（或前 2k 线性预热至 0.15）。

*训练时长与批设置：*单序列训练 30∼60k 迭代；每迭代随机采样 N_{tiles} 个瓦片（默认 8）和每瓦片 M 个像素（默认 1024）。所有方法的训练总步数、分辨率与教师缓存维度一致，确保比较公平。

上述设置在不改变 3DGS/4DGS 渲染复杂度的前提下即可复现本文结果；所有随机种子固定并在附录公开配置与日志，以便复现与对照。

*e) **KITTI Dataset**:* KITTI 系列数据集提供了典型的城市/郊区道路驾驶场景，多传感器同步采集（RGB 相机、GPS/IMU、激光雷达等）以及完整的相机标定。我们在实验中仅使用与本方法相关的视觉与标定信息：逐帧 RGB 图像、相机内参数（焦距、主点、畸变系数）与外参（相机—车辆坐标系的位姿）、以及逐帧时间戳。为避免对额外监督的依赖，我们不使用深度、稠密语义标注、三维检测框或激光雷达。数据预处理方面，我们对图像进行去畸变与尺寸归一化（长边缩放至 800，保持纵横比），

表 I
KITTI-360 上的主结果（插值/外推设置）。粗体为最佳，下划线次优。

Method	PSNR↑	SSIM↑	LPIPS↓	S-PSNR↑	D-PSNR↑	SVR↓
3DGS						
4DGS						
PVG (w/o SD)						
CoDa-4DGS						
SiT-PVG (ours)						

表 II
WAYMO OPEN 上的主结果（插值/外推设置）。

Method	PSNR↑	SSIM↑	LPIPS↓	S-PSNR↑	D-PSNR↑	SVR↓
3DGS						
4DGS						
PVG (w/o SD)						
CoDa-4DGS						
SiT-PVG (ours)						

并将内参按同样比例缩放；仅做亮度/对比度的轻度归一化，不引入几何增强以免影响几何收敛。训练/测试划分采取时间上不重叠的连续片段：训练片段用于拟合 3D/4D 表征与语义蒸馏，测试片段既包含训练视角的已见时刻（插值评测）也包含未见时刻/新视角（外推评测）。对每条序列，我们优先挑选包含“静态大背景（路/楼/天）+ 典型动态体（车/人）”的片段，以覆盖本文关注的动静共存路况。

f) ***Waymo Open Dataset****:* Waymo Open 提供车载多相机（前/后/左右视）与多激光雷达的同步序列、精确的时间戳和相机/雷达外参。为与本文的纯视觉设定一致，我们同样仅使用逐帧 RGB 图像及其相机内外参与时间戳，不使用激光雷达及三维标注。我们沿用与 KITTI 相同的图像预处理与标定缩放策略（去畸变、长边 800、内参同比缩放），并采用时间上不重叠的片段划分；测试同时覆盖已见时刻与未见时刻/新视角。考虑到 Waymo 的视角覆盖更广、动态目标更丰富，我们在采样片段时优先选择车辆/行人密集、交通参与者运动形态多样的路段，以检验模型在复杂交通动态下的稳定性与可扩展性。

g) ***两数据集的共用设定与我们如何使用****:* 无论是 KITTI 还是 Waymo，我们的输入仅为：RGB 图像 I_t、相机内参 $\mathbf{K}_t$、相机外参（或车辆位姿）$\mathbf{E}_t$ 以及时间戳 t。这些信息用于：(*i*) 3DGS/4D 表征的可微光栅化（把 3D/4D 高斯投影到像素平面）；(*ii*) 构造时间一致的可见性与 α-合成权重；(*iii*) 通过 t 与相机位姿定义相邻时刻（默认 Δ=1 帧）用于速度图 $V(\mathbf{p}, t)$ 和时间约束。我们不使用任何对象级标签（如 3D 框、跟踪 ID），也不使用激光雷达/深度作为监督信号。语义信息来自冻结的 2D 教师（LSeg）：对训练帧离线前向得到像素级语义特征图 $F_t = \mathrm{LSeg}(I_t)$，以 FP16 在 $H/2$ 或 $H/4$ 尺度缓存（可选 PCA 压缩到 64/128 维），训练时上采样并经 1×1 线性头 $U(\cdot)$ 对齐维度，分别用于 2D→4D 语义蒸馏与教师—学生特征差异（构造动静证据）。时间相关的训练安排在两数据集上保持一致：单阶段端到端优化，温启动地提升语义速度约束权重 λ_v 与静态度下界 $\rho^\star$，使模型先完成可靠的外观/几何收敛，再逐步写入“静区近零速度、长期稳态”的时间先验。

h) ***实现细节（与设备环境）****:* 所有实验基于 PyTorch，在 **1× NVIDIA RTX 4090 (24GB)** 上运行；更长序列可在 **1× A100 (80GB)** 复现。混合精度采用 FP16/BFloat16，梯度裁剪阈值 1.0。时间相关参数中性初始化（$\boldsymbol{v}$=**0**, l=1, β=1），语义门控与动静融合器线性层零初始化（$g \approx 0.5$, $\delta \approx 0.5$），按§3.6 的学习率与权重调度训练；每迭代在时刻 t 及对称步 $t \pm \Delta$（Δ=1）采样瓦片像素，并依据上一轮静态概率 $\delta(\mathbf{p}, t)$ 做分层采样（静/动各半）以稳定速度损失的梯度。为公平起见，KITTI 与 Waymo 上的训练分辨率、训练步数、教师特征维度与缓存策略保持一致；所有随机种子固定，训练/测试片段划分将随代码一并公开以便复现。

G. 实验设置（*Experimental Setup)*

本节给出用于评估 SiT-PVG 的数据集、评测指标、基线与实现细节，确保结果可复现且与既有方法公平对比。

a) 数据集（Datasets）： **Waymo Open Dataset（WOD）.** WOD 提供同步的多相机与多线激光雷达序列以及精确的时间戳与标定信息。我们选取 *8* 段城市驾驶序列（其中含 *2* 段高度动态场景），主要使用前向的 *3* 个相机视角作为训练/评测视角，分辨率设为 1920×1280；另留出前向的第 *4* 个相机作为新视角用于 NVS 评估。图像做去畸变与亮度/对比度轻度归一化，相机内参与外参采用官方标定并按缩放比例线性调整。为提升初始化稳定性，我们**仅将车载 LiDAR 的稠密点云用于高斯的几何播种（seeding）**，不将 LiDAR 作为训练监督或评测信号。

KITTI Dataset. KITTI 提供立体/多目相机序列与车辆位姿信息。参照自监督动态重建常用协议，我们选择 *4* 条运动丰富的序列，采用左右两个相机视角（1242×375）进行训练与评测；分辨率与标定的处理方式与 WOD 相同。若该序列提供 Velodyne 点云，则**仅用于几何播种**，不参与监督；未提供点云时，使用基于 SfM 的稀疏重投影初始化。两数据集均采用时间不重叠的训练/测试切分，测试既包含训练视角的已见时刻（4D 重建）也包含未见时刻/新视角（NVS）。

b) 评测指标（Metrics）： 我们遵循道路场景新视角合成与 4D 重建的通用做法，报告 **PSNR**（↑）、**SSIM**（↑）与 **LPIPS**（↓）三个全图指标。为单独衡量动静区域的重建质量，进一步给出 *静态区域 PSNR*（S-PSNR）与 *动态区域 PSNR*（D-PSNR）：区域划分来自第 3.3 节提出的*双线索动态掩码*（DEMM），其中教师–学生特征差异提供动态线索，LSeg 语义先验提供静态线索，二者经逻辑回归融合得到像素静态概率 $\delta(\mathbf{p}, t)$ 并阈值化生成掩码。除非另行说明，时间步 $\Delta{=}1$，所有指标均在统一分辨率与相机标定下计算。

c) 基线（Baselines）： 我们选取当前具有代表性的 NeRF 与 Gaussian Splatting 系方法进行对比：**SUDS**、**EmerNeRF**（自监督 NeRF 类）；**3DGS**（静态显式点元）、**4DGS/S3Gaussian**（时间扩展的 GS）、**PVG**（周期振荡参数化）、**StreetGaussian**、**CoDa-4DGS**、**DeSIRE-GS**（语义/动态增强的 GS 类）。其中 *SUDS*、*EmerNeRF*、*PVG*、*CoDa-4DGS*、*DeSIRE-GS* 均可在不使用对象级标注的设定下运行；若某些基线默认依赖对象级监督（如 3D 边界框/跟踪），我们在无标注配置下运行或在表格中单独标注。所有基线统一训练步数、输入分辨率与相机标定；涉及语义特征的基线统一使用 LSeg 作为 2D 教师以剥离语义质量差异。

d) 实现细节（Implementation Details）： *硬件与框架.* 所有实验基于 PyTorch，在 **1× vGPU (48 GB)** 上训练与评测（长序列可在 A100 80GB 等效运行）；使用混合精度（FP16/BFloat16）与梯度裁剪（阈值 1.0）。

初始化与播种. 以 LiDAR（若可用）或 SfM/VO 得到的点云进行高斯播种；时间相关参数中性初始化：$\boldsymbol{v}_i{=}\mathbf{0}$, $l_i{=}1$, $\beta_i{=}1$。语义门控与动静融合器的线性层置零，使 $g_i{\approx}0.5$, $\delta{\approx}0.5$，避免早期过抑制。

优化与调度. 优化器为 Adam（$\beta_1{=}0.9$, $\beta_2{=}0.999$）。学习率与损失权重遵循第 3.6 节：渲染/几何分支初始 $(1\sim2)\times10^{-3}$，语义/时间分支 5×10^{-4}（第 1k 次迭代升至 1×10^{-3}），余弦退火至初值的 $1/10$。权重温启动：λ_v 前 5k 迭代从 0 线性升至 0.5；$\rho^\star$ 前 15k 迭代从 1.0 升至 1.5；$\lambda_\rho{=}0.15$。重建项 $\mathcal{L}_{\text{rgb}}$ 为 L_1 与 SSIM 的加权组合；语义蒸馏 $\mathcal{L}_{\text{SD}}$ 采用像素级 L_1；时间约束包含语义速度约束 $\mathcal{L}_v$ 与静态寿命先验 $\mathcal{L}_\rho$，并辅以轻量 ℓ_2 正则 $\mathcal{L}_{\text{reg}}$。

教师特征与批采样. 训练帧离线前向 LSeg 得到 F_t（FP16，$H/2$ 或 $H/4$，可选 PCA 压至 64/128 维），训练时上采样并经 $1{\times}1$ 对齐头参与蒸馏与差异计算。每次迭代在时刻 t 及对称步 $t\pm\Delta$（$\Delta{=}1$）进行瓦片化像素采样；为缓解动静不平衡，按上一轮静态概率 $\delta(\mathbf{p}, t)$ 做分层采样（静/动各半），稳定 $\mathcal{L}_v$ 的梯度。

公平性约束. 基线与我方法在相同的训练/测试切分、分辨率、训练步数与相机标定下运行；涉及语义的比较统一使用相同教师与特征维度；若某方法天然依赖对象级监督，则在表格中明确标注其监督级别与使用的外部先验。

H. 实验设置（Experimental Setup）

本节给出用于评估 SiT-PVG 的数据集、评测指标、对比基线与实现细节；其中若干设定分别参照既有工作的常用协议（文末括注注明来源，正式稿中再补充引用）。

a) 数据集（Datasets）： **Waymo Open Dataset（WOD）.** 我们参照 *OmniRe* 的场景与相机使用方式，选取 8 段城市道路序列（其中包含 2 段高度动态场景），主要使用前向三个相机作为训练/评测视角（分辨率 1920×1280），并保留前向第 *4* 个相机作为新视角用于 NVS 评估。图像做去畸变与亮度/对比度轻度归一化，相机内参与外参采用官方标定并按缩放比例线性调整。我们只将车载 LiDAR 的稠密点云用于高斯播种（几何初始化），不将其作为训练监督或评测信号（做法与若干基于高斯的道路重建工作一致）。 **KITTI Dataset.** 我们参照 *SUDS* 的 *KITTI* 设定，选择 4 条运动显著的序列，采用左右两个相机（分辨率 1242×375）进行训练与评测；预处理与标定缩放与 WOD 相同。若该序列提供 Velodyne 点云，则仅用于高斯播种；未提供点云时，使用

基于 SfM/VO 的稀疏重投影初始化。两数据集均采用时间不重叠的训练/测试切分：测试同时包含训练视角的已见时刻（4D 重建）与未见时刻/新视角（NVS），与道路场景 4D/NVS 常用评测协议一致（做法与 CoDa-4DGS 的分任务评测口径一致）。

b) 评测指标（Metrics）： 我们报告三项通用图像质量指标：**PSNR**（↑）、**SSIM**（↑）与 **LPIPS**（↓）。为单独衡量动静区域的重建质量，进一步给出 静态区域 *PSNR*（S-PSNR）与 动态区域 *PSNR*（D-PSNR），其区域划分来自第 3.3 节提出的双线索动态掩码（*DEMM*）：以教师–学生特征差异作为动态线索、以 LSeg 文本对齐语义作为静态先验，经逻辑回归融合得到像素静态概率 $\delta(\mathbf{p}, t)$ 并阈值化生成掩码。该“动态区域专用指标”的呈现方式与近期 4D 高斯语义增强方法一致；除非另行说明，时间步 Δ=1，所有指标均在统一分辨率与相机标定下计算。

c) 基线（Baselines）： 我们覆盖两大类强基线并统一其运行配置：*NeRF* 系——SUDS、EmerNeRF（均为自监督/弱监督设定）；*Gaussian Splatting* 系——3DGS（静态显式点元）、4DGS/S3Gaussian（时间扩展）、PVG（周期振荡参数化）、Street-Gaussian、CoDa-4DGS、DeSIRE-GS（语义/动态增强）。其中 SUDS、EmerNeRF、PVG、CoDa-4DGS、DeSIRE-GS 可在无对象级标注下运行；若某方法默认依赖对象级监督（如 3D 边界框/跟踪），我们在无标注配置下运行或在主表中单列标注其监督级别。为剥离语义质量差异，涉及语义的基线统一采用 LSeg 作为 2D 教师；训练步数、输入分辨率与相机标定等均与我方法保持一致（这一“统一协议”做法与 PVG/CoDa-4DGS 的对比设置一致）。

d) 实现细节（Implementation Details）： 所有实验基于 PyTorch，在 **1× vGPU 48GB** 上以混合精度（FP16/BFloat16）运行；每次迭代从时刻 t 及对称步 $t \pm \Delta$（Δ=1）进行瓦片化像素采样，并依据上一轮静态概率 $\delta(\mathbf{p}, t)$ 做分层采样（静/动各半）以稳定速度损失梯度；训练帧离线前向 LSeg 得到教师特征 F_t（FP16，$H/2$ 或 $H/4$，可选 PCA 压至 64/128 维），训练时上采样并经 1×1 对齐头参与蒸馏与差异计算；高斯播种由 LiDAR（若可用）或 SfM/VO 稀疏点初始化，时间相关参数中性初始化 $\boldsymbol{v}_i$=**0**, l_i=1, β_i=1，语义门控与动静融合器线性层置零使 $g_i \approx 0.5$, $\delta \approx 0.5$；优化器使用 Adam（β_1=0.9, β_2=0.999），学习率与损失组合参照 *PVG* 的实现口径并结合我方法的时间约束做温启动：渲染/几何分支初始 $(1 \sim 2) \times 10^{-3}$，语义/时间分支 5×10^{-4}（第 1k 次迭代升至 1×10^{-3}），余弦退火至初值的 1/10；λ_v 在前 5k 迭代从 0 线性升至 0.5，$\rho^\star$ 在前 15k 迭代从 1.0 升至 1.5，λ_ρ=0.15，并施加轻量 ℓ_2 正则以抑制语义向量与速度基过大波动；训练/测试切分、步数、分辨率与相机标定在所有方法间统一（这一公平性约束与 CoDa-4DGS 的比较协议一致）。

括注（将来补充引用）：OmniRe（WOD 的序列/相机与 NVS 评估设置）、SUDS（KITTI 的序列选择与左右相机分辨率）、PVG（优化器与损失权重的口径与温启动思路）、CoDa-4DGS（4D/NVS 双任务与动态区专用指标的呈现方式）。

e) 实现细节（Implementation Details）： 所有实验基于 **PyTorch** 实现，默认在 **1× NVIDIA vGPU (48GB)** 环境下运行。我们使用官方 LiDAR 点云进行高斯初始分布拟合以稳定优化过程。优化器采用 **Adam**，初始学习率为 2×10^{-3}，采用余弦退火策略逐步下降至 1×10^{-4}。批大小为 2，梯度裁剪阈值设为 1.0。时间相关参数中性初始化（v_i=**0**, l_i=1, β_i=1），语义门控与融合器线性层初始化为零（$g_i \approx 0.5$, $\delta \approx 0.5$），以避免早期过抑制动态。训练分两阶段进行：第一阶段在 1/4 分辨率下预热以稳定几何与外观，第二阶段逐步升至全分辨率联合优化语义与时间约束。损失函数权重遵循§3.6 中的温启动策略：λ_{SD}=1.0，λ_v 在前 5K 次迭代从 0 升至 0.5，$\rho^\star$ 在前 15K 次迭代从 1.0 升至 1.5，λ_ρ=0.15。训练收敛后，我们对 Waymo 与 KITTI 上的测试序列分别执行 4D 重建与新视角渲染评估。

I. 比较结果（*Quantitative & Qualitative Results*）

本节从两方面报告 SiT-PVG 的实验结果：定量评估（4D 重建与新视角合成的客观指标）与定性评估（可视化对比与误差分析）。评测协议、数据与训练设置见 §**??**。

a) 定量评估（Quantitative Evaluation）： 我们在 **Waymo Open** 与 **KITTI** 上分别报告全图 PSNR/SSIM/LPIPS，以及基于动静掩码的 静态区域 *PSNR*（*S-PSNR*）与 动态区域 *PSNR*（*D-PSNR*）。同时，为衡量静区伪运动，给出 静态速度残差（SVR，详见 §3.4 中的 $V(\mathbf{p}, t)$ 定义）。表 III 与表 IV 为主结果；表 V 给出消融研究，包括去除 2D→4D 语义蒸馏（w/o SD）、去除语义速度约束（w/o SVC）与去除静态寿命先验（w/o SLP）等变体。

为便于后续填表，以下表格先行提供**占位结构**（列名/方法名已对齐常见写法）。填入数值后，摘要性描述可据实际结果微调（例如“在 S-PSNR 与 SVR 上领先”“D-PSNR 与 4D 基线相当”等）。

表 III
WAYMO OPEN 上的主结果（4D 重建与新视角合成）。↑越大越好，↓越小越好。PSNR/SSIM/LPIPS 为全图指标；S-PSNR, D-PSNR 分别在静/动态区域统计；SVR 衡量静区伪运动。

方法	PSNR↑	SSIM↑	LPIPS↓	S-PSNR↑	D-PSNR↑	SVR↓
3DGS						
4DGS						
PVG (w/o SD)						
SUDS						
EmerNeRF						
Street-Gaussian						
CoDa-4DGS						
DeSiRe-GS						
SiT-PVG (Ours)						

表 IV
KITTI 上的主结果（与表 III 指标一致）。

方法	PSNR↑	SSIM↑	LPIPS↓	S-PSNR↑	D-PSNR↑	SVR↓
3DGS						
4DGS						
PVG (w/o SD)						
SUDS						
EmerNeRF						
Street-Gaussian						
CoDa-4DGS						
DeSiRe-GS						
SiT-PVG (Ours)						

表 V
消融研究（以 WAYMO 为例）。从 PVG 基线出发，逐步引入 SD/SVC/SLP。

变体	S-PSNR↑	D-PSNR↑	SVR↓	LPIPS↓
PVG（基线）				
+ SD				
+ SD + SVC				
+ SD + SVC + SLP（Ours）				

占位图 A：静态速度热图（越暗越静）。
(建议：每列同一帧，行分别为 GT、PVG、CoDa-4DGS、Ours；用 $V(\mathbf{p},t)$ 伪彩上色)

图 1. 静态速度热图对比。SiT-PVG 在路面与建筑区域显著更暗，遮挡边界处抖动减小。

撰写提示（放表后可据实修改）：若观察到 S-PSNR 与 SVR 明显改善，可写为：“在两套基准上，SiT-PVG 在 S-PSNR 与 SVR 上持续优于 3DGS/4DGS/PVG 等基线，表明静态背景伪运动得到有效抑制；同时在 D-PSNR 上与 4D 基线相当，说明对前景真实运动的保留没有显著损害。”

b) 定性评估（Qualitative Evaluation）： 我们从三类可视化考察方法行为：（i）静区速度热图；（ii）新视角/新时刻渲染；（iii）局部细节与边界。为便于先排版，以下图示使用占位框，待导出图片后替换为 `\includegraphics`。

制作建议： 1) 速度热图：按式（100）生成 $V(\mathbf{p},t)$，归一化至 $[0,1]$ 并用同一色标显示；静区应更暗。 2) 新视角/新时刻：固定一个未见视角，展示 $t-\Delta$, t, $t+\Delta$ 三帧；如有

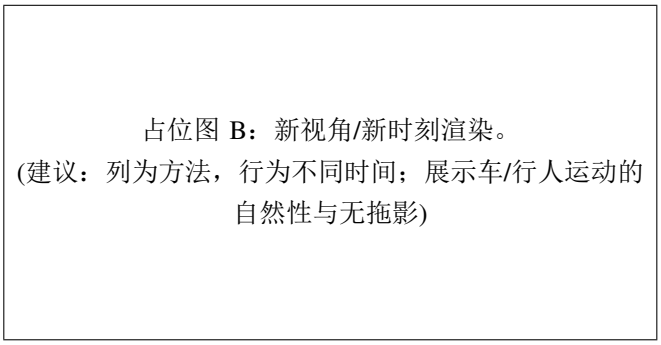

图 2. 新视角与时间插值的可视化。SiT-PVG 在前景轨迹上保持连贯且无明显拖影。

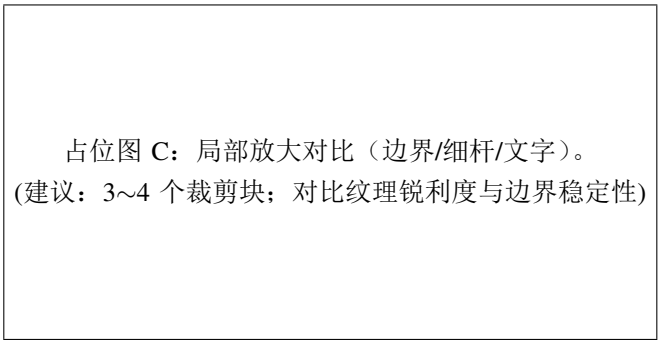

图 3. 局部细节对比。我们方法在细长结构与高对比边界处更稳定，纹理保真度更高。

GT，可附上真值。 3) 局部裁剪：选择（车灯、轮廓、路牌边缘、电线等）对比易漂/易抖区域，统一 4× 放大。

c) 误差与局限（Failure Cases & Discussion）: 在强遮挡或极远距离下，教师特征与可见性会同步变弱，动静掩码可能误检，导致个别帧出现轻微残影；在极长序列（>数千帧）时，若无额外闭环约束，仍可能累积微小漂移。对于这些情况，引入更强的跨时序一致性（如 TCC）或多教师融合可进一步缓解（可在附录给出对比）。

IV. 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-to-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），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显式建模场景。每个基元携带位置、各向异性尺度与旋转、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

$$G_i(\boldsymbol{x}) = \exp\Big(-\tfrac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\Big). \tag{88}$$

其中，$\boldsymbol{\mu}_i$ 为基元在三维空间中的位置；$\boldsymbol{\Sigma}_i = \mathbf{R}_i\mathbf{S}_i\mathbf{S}_i^\top\mathbf{R}_i^\top$ 为协方差矩阵（$\mathbf{R}_i$ 为旋转，$\mathbf{S}_i$ 为尺度）。将三维高斯投影到像平面得到 2D 高斯，其投影协方差近似为

$$\boldsymbol{\Sigma}_i' = \mathbf{J}\,\mathbf{W}\,\boldsymbol{\Sigma}_i\,\mathbf{W}^\top\mathbf{J}^\top, \tag{89}$$

其中 $\mathbf{W}$ 为世界到相机的外参变换，$\mathbf{J}$ 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：

$$C = \sum_{i=1}^{N} T_i\,\alpha_i\,c_i, \qquad T_i = \prod_{j<i}(1-\alpha_j), \tag{90}$$

其中 α_i 由不透明度与投影 2D 协方差在该像素的覆盖共同决定，c_i 为外观。该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 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/幅度向量 $\boldsymbol{v}$、寿命尺度 β，定义

$$\tilde{\boldsymbol{\mu}}(t) = \boldsymbol{\mu} + \frac{l}{2\pi}\sin\Big(2\pi\frac{t-\tau}{l}\Big)\,\boldsymbol{v}, \qquad \tilde{o}(t) = o\cdot\exp\Big(-\tfrac{1}{2}\,(t-\tau)^2\,\beta^{-2}\Big). \tag{91}$$

时刻 t 的点元状态为 $H(t) = \{\tilde{\boldsymbol{\mu}}(t), \mathbf{q}, \mathbf{s}, \tilde{o}(t), \mathbf{c}\}$（其中 $\mathbf{q}$ 为旋转参数、$\mathbf{s}$ 为尺度参数、$\mathbf{c}$ 为外观），整幅图像按

$$\hat{I}_t = \mathrm{Render}\big(\{H_i(t)\}_{i=1}^{N};\,\mathbf{K}_t,\,\mathbf{E}_t\big) \tag{92}$$

渲染，其中 $\mathbf{K}_t, \mathbf{E}_t$ 分别为相机内参与外参。为衡量点元“静态度”，定义

$$\rho = \beta/l, \tag{93}$$

ρ 越大表示寿命相对周期更长、越趋近静止；当 $\boldsymbol{v} = \mathbf{0}$ 且 $\rho \to \infty$ 时，PVG 退化回 3DGS。以上设计使静态/动态以统一形式出现，仅通过 $\{\boldsymbol{v}, \beta, l, \tau\}$ 的取值区分；PVG 以最小改动继承 3DGS 的高效性，同时补足动态建模与可编辑性。

B. 2D-to-4D 语义蒸馏

仅凭重建损失难以稳定区分“相机运动”与“真实世界运动”。为增强模型对场景语义的理解，我们将 2D 基础模型的稠密语义特征迁移至 4D 高斯表征，使每个点元学习到连续可度量的语义向量，便于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。

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

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

学生端为每个高斯基元赋予可学习语义向量 $f_{\mathrm{sem},i}$。与 RGB 渲染一致，像素 $\mathbf{p}$ 处的学生语义按可见性权重加权聚合；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

$$F_s(\mathbf{p},t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\, f_{\mathrm{sem},i}. \tag{94}$$

为与教师通道数对齐，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：

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

采用像素级 L_1 蒸馏使学生贴近教师，构建蒸馏损失

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

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}，形成可随 PVG 形变在时间上搬运的 4D 语义表征。

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 $\tilde{F}_s$。本节据此构造由两类证据共同约束的帧级动静掩码：*教师–学生特征差异*与*语义先验*。前者反映像素处的语义不一致，后者提供先验上应静止的区域指示，二者经轻量融合得到稳定的静态掩码 M_{stat} 与动态掩码 M_{dyn}。

a) (a) 教师–学生特征差异: 当像素对应静态背景时，多帧观察指向同一世界点，学生的语义渲染应与教师一致；当像素位于动态对象或遮挡边界时，更易产生偏差。定义像素级余弦不相似度：

$$D(\mathbf{p},t) = 1 - \cos\big(\tilde{F}_s(\mathbf{p},t),\, F_t(\mathbf{p})\big). \tag{95}$$

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) 语义先验与融合: 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：

$$M_{\mathrm{sem}}(\mathbf{p},t) = \sum_{k\in\mathcal{C}_{\mathrm{stat}}} S_k(\mathbf{p},t). \tag{96}$$

为避免手工权重，用两输入一输出的逻辑回归融合两条证据并输出静态概率：

$$\delta(\mathbf{p},t) = \sigma\big(a\cdot(1-D(\mathbf{p},t)) + b\cdot M_{\mathrm{sem}}(\mathbf{p},t) + c\big), \tag{97}$$

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，δ 作为*软静态权重*直接用于§3.4 的时间约束；当导出或评测掩码时，再以固定阈值二值化：

$$M_{\mathrm{stat}}(\mathbf{p},t) = \mathbf{1}\big(\delta(\mathbf{p},t) > \tau_s\big), \qquad \tau_s = 0.5, \tag{98}$$

并令 $M_{\mathrm{dyn}}(\mathbf{p},t) = 1 - M_{\mathrm{stat}}(\mathbf{p},t)$。为降低误检，采用保守融合策略（静态取交、动态取并），此处不再展开额外公式。

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 $f_{\mathrm{sem},i}$ 产生速度门控并直接作用于 PVG 的速度基向量：

$$g_i = \sigma(\mathbf{w}_g^\top f_{\mathrm{sem},i} + b_g) \in (0,1), \qquad \boldsymbol{v}_i^{\mathrm{eff}} = g_i\,\boldsymbol{v}_i, \tag{99}$$

随后以 $\boldsymbol{v}_i^{\mathrm{eff}}$ 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：

$$V(\mathbf{p},t) = \sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,\big\|\Pi\big(\boldsymbol{\mu}_i(t+\Delta)\big) - \Pi\big(\boldsymbol{\mu}_i(t-\Delta)\big)\big\|_1. \tag{100}$$

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：

$$\mathcal{L}_v = \mathrm{mean}_{\mathbf{p}\in\Omega}\big(\delta(\mathbf{p},t)\cdot V(\mathbf{p},t)\big). \tag{101}$$

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

$$\rho_i = \beta_i / l_i, \tag{102}$$

并用 w_i^{stat} 对静态倾向更高的点元施加*下界约束*：

$$\mathcal{L}_\rho = \sum_i w_i^{\mathrm{stat}}\,\max\big(0,\ \rho^\star - \rho_i\big), \tag{103}$$

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。若在几何尚未收敛时过早强化时间约束，容易把真实运动过度抑制或引发训练抖动。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。本文的重建项 $\mathcal{L}_{\mathrm{rgb}}$ 由像素级 L_1 与 SSIM 的加权组合构成，用于驱动几何与外观收敛而不额外引入假设。

a) （可选）辅助几何监督：深度正则*:* 为进一步稳定几何，我们以与颜色一致的权重渲染期望深度：

$$\hat{D}(\mathbf{p},t)=\sum_{i\in\mathcal{V}(\mathbf{p},t)} w_i(\mathbf{p},t)\,d_i(t),$$

其中 $d_i(t)$ 为 $\boldsymbol{\mu}_i(t)$ 在相机坐标系下的 z 向深度。可使用稀疏 LiDAR 对齐损失 $\mathcal{L}_{\mathrm{depth}}^{\mathrm{lidar}}$（仅在有激光点的像素上计算）或多视角自监督一致性 $\mathcal{L}_{\mathrm{depth}}^{\mathrm{mv}}$（将深度投影到邻帧比较），并配合边缘保持平滑 $\mathcal{L}_{\mathrm{sm}}$。该分支为可选，不改变前向渲染逻辑。

b) 总体目标与权重调度*:* 总体目标为

$$\mathcal{L}=\mathcal{L}_{\mathrm{rgb}}+\lambda_{\mathrm{SD}}\mathcal{L}_{\mathrm{SD}}+\lambda_v\mathcal{L}_v+\lambda_\rho\mathcal{L}_\rho+\lambda_{\mathrm{depth}}\mathcal{L}_{\mathrm{depth}}+\lambda_{\mathrm{sm}}\mathcal{L}_{\mathrm{sm}}+\lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}},\tag{104}$$

其中 $\mathcal{L}_{\mathrm{SD}}$ 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；$\mathcal{L}_{\mathrm{depth}}$ 表示选用的深度对齐项（LiDAR 或多视角一致性），$\mathcal{L}_{\mathrm{sm}}$ 为深度平滑，$\mathcal{L}_{\mathrm{reg}}$ 为轻量正则。为避免早期抑制动态与造成不稳定，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；$\lambda_\rho\equiv 0.15$；$\rho^\star$ 在前 15k 次迭代由 1.0 线性升至 1.5；深度分支若启用，$\lambda_{\mathrm{depth}}\in[0.1,0.5]$、$\lambda_{\mathrm{sm}}\in[1\times10^{-3},5\times10^{-3}]$ 并同样采用 3–5k 的线性升权。

为抑制语义向量与速度基的过大波动，引入 ℓ_2 正则：

$$\mathcal{L}_{\mathrm{reg}}=\lambda_f\sum_i\|f_{\mathrm{sem},i}\|_2^2+\lambda_v^{\ell_2}\sum_i\|\boldsymbol{v}_i\|_2^2,\tag{105}$$

默认 λ_f=1×10^{-4}，$\lambda_v^{\ell_2}$=5×10^{-5}。上述设置在保持高斯渲染效率的同时，逐步写入语义与时间先验，从而减少静区伪运动并保留前景真实运动。