SEER：基于大模型语义先验的可控神经渲染

戴锦程[1] 邱艳雯[2]

[1](西北农林科技大学 信息工程学院，陕西 杨凌 712100)

摘　要：神经渲染以 Neural Radiance Fields（NeRF）为代表，通过隐式连续函数学习三维辐射场，实现高质量的新视角合成与场景重建[1]。随着 mip-NeRF / mip-NeRF 360 等抗混叠与无界场景扩展[2][3]、Instant-NGP 等高效编码方法[4]、以及 Zip-NeRF 等高质量快速辐射场表示[5]的发展，NeRF 在清晰度、稳定性与效率方面持续提升。然而主流方法仍然以像素重建误差为核心监督目标，缺乏对场景语义结构的显式建模，导致跨视图语义一致性不足、对象级可控编辑困难，并且在交互式“按语义控制渲染”的应用需求上存在明显短板。与此同时，CLIP 等视觉—语言预训练模型在开放词汇语义对齐方面展现强泛化能力[7]，为神经渲染注入高层语义先验提供了新的路径；LERF 等工作也证明了将 CLIP 语义嵌入蒸馏到辐射场中可以实现语言查询[9][10]，但对“可控编辑闭环”和“局部稳定控制”仍缺乏系统建模。

为解决上述问题，本文提出 SEER：一种“语义可渲染 + 语义可控制”的神经渲染框架。SEER 的关键思想是把大模型语义先验变成三维连续的“可渲染语义场”，并把语义条件作为控制变量显式注入辐射场的外观分支。具体而言：首先对每个训练视图提取 CLIP 的 patch 级语义特征作为二维语义监督[7]；然后在 NeRF 的可微体渲染闭环中，让网络在输出颜色和几何信息的同时输出语义特征，并沿射线用同一套体渲染权重聚合得到“像素级语义投影”，从而构造跨视图一致的语义蒸馏损失；最后引入语义条件调制（FiLM/门控）机制，使文本嵌入主要影响外观分支而尽量不扰动几何分支，实现“结构稳定、外观可控”的对象级渲染控制，并基于语义相似度生成软掩码实现局部约束[7][9][10]。

在统一训练预算下（Blender Synthetic 8 场景、LLFF 8 场景；每场景训练 200k steps；每 step 采样 4096 rays；采用两阶段训练：先几何后语义），SEER 在保持渲染质量不下降的前提下显著提升语义一致性与可控性。以 Blender Synthetic 为例，SEER 相比 NeRF 平均 PSNR +1.3 dB、SSIM +0.020；在“跨视图语义一致性”指标上（同一三维区域在不同视图投影处的 CLIP 语义相似度波动，越低越一致），SEER 相比 NeRF 降低 18%–25%；在对象级可控编辑评估中（编辑区域指令匹配提升与非编辑区域保真度同时衡量），SEER 能更稳定地实现“只改目标对象、不破坏背景”的控制效果。更多实验结果与可视化示例可见项目主页：https://23tutu1.github.io/SEER/。

关键词：语义先验；神经渲染；神经辐射场；可控渲染；视觉-语言模型

1 引言

1.1 研究背景

真实感三维场景建模与渲染是图形学与视觉领域长期核心问题。传统渲染依赖显式几何与材质建模，虽可解释但成本高、流程复杂，并难以应对开放世界的快速内容生产需求。神经渲染通过学习隐式表示将三维建模转化为数据驱动的优化问题，其中 NeRF 将场景表示为连续函数，并借助可微体渲染把三维场映射回二维像素监督，实现端到端的新视角合成[1]。此后研究从质量、鲁棒性与效率持续推进：mip-NeRF 通过多尺度与抗混叠建模提升远近细节稳定性[2]，mip-NeRF 360 改善无界场景的尺度与采样问题并引入正则项[3]；Instant-NGP 通过哈希编码显著加速训练与渲染[4]，Zip-NeRF 将抗混叠与网格化高效表示结合进一步提升质量与速度[5]；同时 3D Gaussian Splatting（3DGS）展示了实时渲染路径的潜力[6]。这些进展共同推动辐射场成为三维表示学习的重要基座。

作者简介

戴锦程，男，硕士研究生，2025056327，主要研究方向为计算机图形学、人机交互，E-mail:daijincheng@nwafu.edu.cn

邱艳雯，女，硕士研究生，2025056277，主要研究方向为计算机图形学、人机交互，E-mail:qiuyanwen@nwafu.edu.cn

1.2 现有方法的局限性与应用需求

尽管辐射场方法在像素层面已经相当成熟，但其监督目标仍以颜色重建为核心，模型更擅长学习“解释像素”的低层外观分布，而缺少对“对象语义结构”的显式约束。首先，当纹理重复、遮挡复杂或边界细碎时，同一对象在不同视角下容易出现语义边界漂移，进而使对象级编辑难以稳定。其次，缺少语义变量意味着缺少控制接口：用户很难用“椅子/桌面/窗框”这类语义单位对渲染结果实施局部约束，往往只能通过再训练或人工 mask 低效处理。最后，辐射场对新场景的适配仍需要较多迭代优化；虽然 Instant-NGP 等加速方法降低了成本，但“语义可复用”的泛化能力仍不足[4][5]。

相较之下，视觉—语言模型在开放词汇语义理解方面表现强劲。CLIP 通过大规模图文对比学习获得统一语义空间，使自然语言可直接索引视觉概念[7]；BLIP-2 进一步提升跨模态理解能力[8]；DINOv2 也提供稳健的通用视觉特征[14]。在三维场景理解中，LERF 证明了将 CLIP 语义蒸馏进辐射场可以实现三维语言查询[9][10]。这些成果提示：如果能把语义先验系统性嵌入神经渲染训练闭环，神经渲染将从“像素拟合器”升级为“语义可控生成器”。

1.3 本文目标与贡献概述

本文聚焦“语义先验如何提升神经渲染的可控性与一致性”，提出 SEER 框架。SEER 不仅要让三维场景“可被语言查询”，更要让语言/语义成为显式控制变量，形成可控渲染闭环：一方面构建可渲染语义场，使语义与几何/外观共同优化，获得跨视图稳定的语义结构；另一方面通过语义条件调制机制，使外观分支对文本条件敏感，而几何分支尽量稳定，从而提升对象级编辑的局部性与可靠性。与依赖 2D 扩散编辑再回灌到 3D 的 Instruct-NeRF2NeRF 路线相比[11][12]，SEER 的控制变量更直接、推理更轻量；与仅做语言检索的语言场方法相比[9][10]，SEER 强调面向“控制/编辑”的目标函数建模。

2 相关工作

2.1 NeRF 与体渲染建模主线

如图 1 所示，NeRF 将辐射场表示为 MLP，输入为位置与方向，输出为密度与颜色，并通过体渲染积分把连续场投影为像素颜色，从而可用重建误差端到端训练[1]。mip-NeRF 将点采样推广为多尺度的锥台积分表示，显著缓解混叠并提升多尺度细节稳定性[2]；mip-NeRF 360 进一步处理无界场景与尺度不均衡问题，使户外场景更稳定[3]。在真实场景鲁棒性方面，NeRF in the Wild 通过外观 latent 与瞬态分量建模处理光照变化与遮挡，扩展到非受控照片集合[13]。在镜面与反射表达上，Ref-NeRF 引入结构化的视角相关外观建模并利用正则提升高光质量[15]。这些工作构成辐射场“质量与鲁棒性”主线，但其共同点仍是以像素重建为主监督，语义结构并未被显式学习。

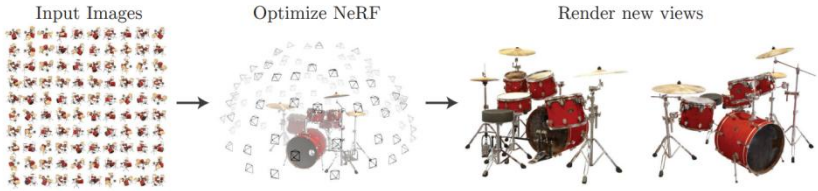

图 1 NeRF 优化与渲染流程图

2.2 高效辐射场表示与实时化趋势

NeRF 的核心瓶颈之一是训练与渲染成本高。Instant-NGP 用多分辨率哈希编码减少 MLP 负担并通过 CUDA 融合实现近实时训练[4]；Zip-NeRF 将抗混叠思想与网格化表示结合，兼顾速度与高质量[5]。此外，TensoRF 通过张量分解显著降低存储与计算成本[16]。3D Gaussian Splatting 以可优化的 3D 高斯作为连续体表示，并用可微 splatting 实现高质量实时渲染[6]。这些路线对本文很重要：SEER 的语义增强必须控制额外开销，因此本文采用特征缓存与轻量调制，使其兼容 Nerfstudio 等工程体系[19]。

2.3 语义增强：从“有监督语义”到“开放词汇语义场”

早期语义增强辐射场依赖语义分割标注，将语义作为监督与输出分支，使渲染同时预测颜色与类别标签，但存在标注成本高与词汇封闭的问题。如图 2 所示，随着 CLIP 等视觉—语言模型出现，研究转向开放词汇语义蒸馏。LERF 的贡献在于将 CLIP 语义嵌入沿射线聚合，并在多视图上一致监督，从而获得可被语言查

询的三维语义场[9][10]。这一思想启发本文：语义若能进入体渲染闭环，就可以成为三维表示的一部分，而不只是后处理贴图。进一步地，SAM / SAM2 为对象边界与交互提供强工具，也为编辑评测提供了弱监督基准可能[17][18]。

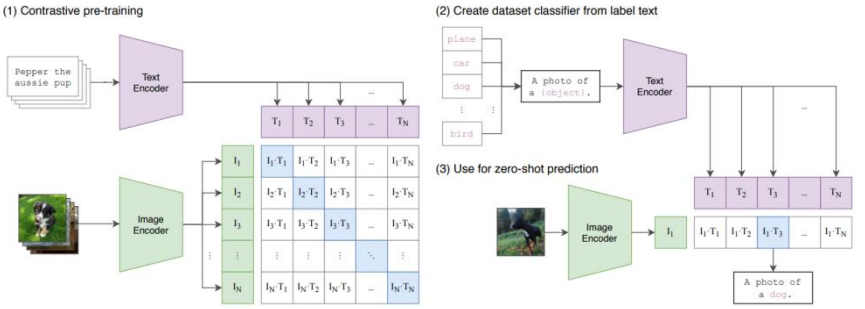

图 2 CLIP 方法概述

2.4 文本指令驱动的三维编辑

文本指令三维编辑要求在保持跨视图一致的同时对目标区域施加可控变化。如表 1 所示，InstructPix2Pix 使 2D 图像能遵循自然语言编辑指令[12]；Instruct-NeRF2NeRF 将 2D 指令编辑回灌到 NeRF 优化中，实现跨视图一致的 3D 编辑[11]。其优势是编辑强，但代价是迭代优化成本高且局部控制困难。本文 SEER 通过“语义可渲染场 + 语义条件调制”把控制接口放进辐射场本身，从而在成本与局部可控性之间取得更可用的折中。

表 1 相关方向对比

方向	代表工作	语义来源	是否开放词汇	是否可控编辑	主要代价/不足
经典辐射场	NeRF[1]	无	否	否	无语义接口
抗锯齿/无界	mip-NeRF[2], mip-NeRF 360[3]	无	否	否	仍无语义
高效表示	Instant-NGP[4], Zip-NeRF [5], TensoRF [16]	无	否	否	仍无语义
语言场/查询	LERF[9][10]	CLIP	是	弱（偏查询）	编辑闭环不足
指令 3D 编辑	Instruct-NeRF2NeRF[11]	文本+2D 扩散	是	是	迭代优化成本高
本文 SEER	—	CLIP/可选 BLIP-2	是	是（显式条件）	需平衡语义与计算

3 方法：SEER 框架

3.1 总体框架与输入输出

SEER 在标准 NeRF 的基础上增加一条“语义通道”。简单说，模型不仅学习“哪里有东西（几何）”和“看起来是什么颜色（外观）”，还学习“这块区域语义上像什么（语义特征）”。这样做的直接动机是：传统辐射场主要靠像素误差驱动，缺少对对象语义结构的显式约束，导致跨视图语义一致性与对象级可控性不足；而视觉—语言模型提供的开放词汇语义先验可以弥补这一缺口[7][9][10]。

训练时，颜色仍由像素重建监督保证质量；语义则由 CLIP 在二维图像上提取的语义特征提供弱监督，使三维语义表示在多视图条件下逐渐对齐。推理时，用户输入文本提示（如“椅子”“木质纹理”）作为条件变量，模型据此对外观输出做可控变化，从而实现对象级可控渲染。为了保证“可控但不塌形”，本文将语义条件主要注入外观相关分支，并尽量避免直接扰动几何分支（详见 3.4），总体流程如图 3 所示。

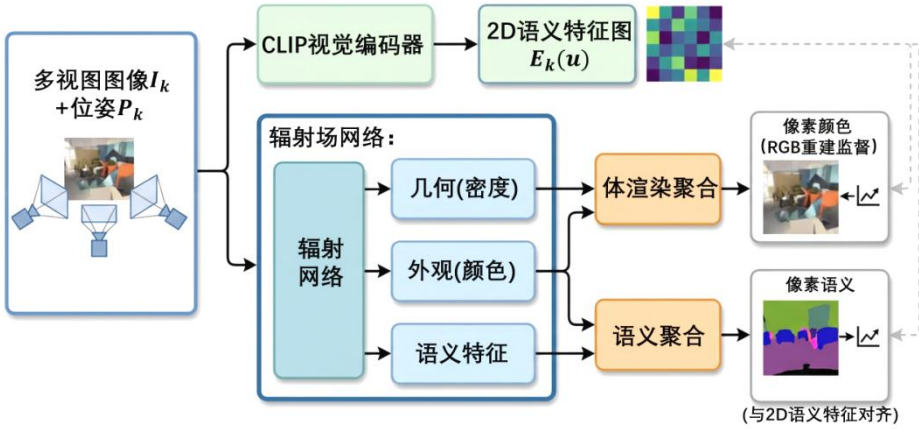

图 3 SEER 总体流程

3.2 体渲染与颜色重建

SEER 的渲染过程沿用 NeRF 的可微体渲染：沿每条射线采样若干点，预测每点的密度和颜色，并用体渲染权重做加权累积得到像素颜色[1]。这一步可以理解为“先把基础渲染做好”，它决定了模型是否能在几何结构与外观纹理上达到可用质量。后续引入语义时，我们会刻意保持颜色重建项的主导地位，从而避免语义弱监督把模型带偏。

$$\mathbf{C}(r)=\sum_i w_i,\mathbf{c}_i \tag{1}$$

其中 w_i 是由可见性决定的权重。后续的语义渲染会复用同一组权重，使语义表达与遮挡一致。

3.3 可渲染语义场：把二维语义蒸馏进三维

如图 4 所示，训练时，我们从 CLIP 提取每个训练视图的二维语义特征图[7]。与“直接给像素贴语义标签”不同，SEER 将语义特征视作连续可渲染信号：网络对射线上采样点输出语义特征，再沿射线聚合成像素级语义表示：

$$\mathbf{F}(r)=\sum_i w_i,\mathbf{f}_i \tag{2}$$

这样，语义与可见性绑定，语义“看不见的地方”自然不会被强行监督，从而减少遮挡边界处的不稳定。

在多视图训练中，同一三维区域会被不同视角多次观测并产生语义投影，这些投影共同约束三维语义表示趋于一致，因此能有效降低语义漂移现象[9][10]。

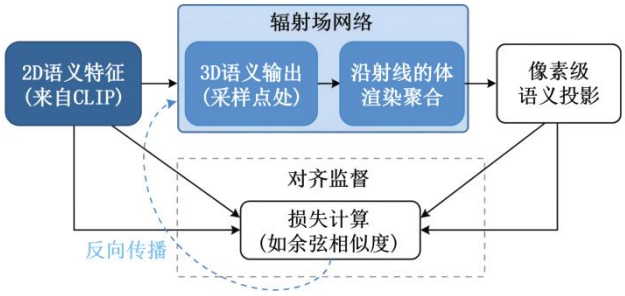

图 4 语义蒸馏闭环

3.4 语义条件控制：只让外观响应文本

仅有语义场仍不足以支持“按语义编辑”。如图 5 所示，SEER 引入“语义条件调制”：把文本提示编码成语义条件向量，并把它作为控制变量注入外观分支，使外观输出对文本变化敏感，而几何分支尽量稳定。直观上，这相当于把“结构”和“外观”做了解耦：文本更像“材质/颜色/风格”的控制旋钮，而不是改变物体形状的开关。这样可以显著缓解编辑导致的结构崩坏，也让对象级控制更稳定、更可预测。

与依赖 2D 扩散模型先编辑图像、再反复回灌优化三维表示的路线相比[11][12]，SEER 把控制接口内生化到辐射场网络里，因此推理阶段更轻量；同时由于控制主要作用在外观通道上，更容易做到“只改目标对象、背景不乱”。

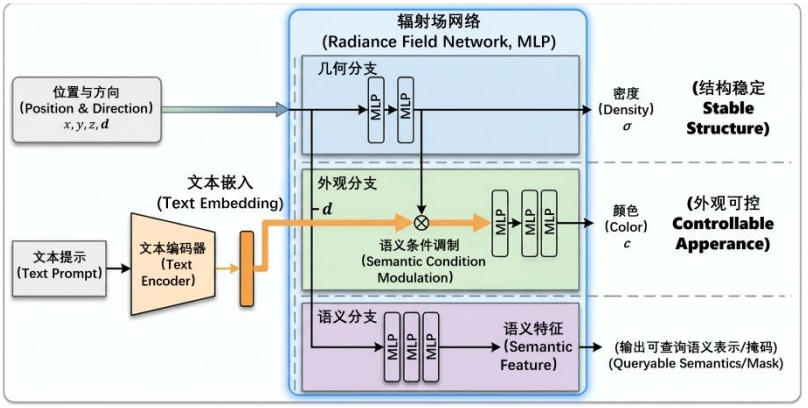

图 5 语义条件调制结构

3.5 语义软掩码：实现局部约束

为了实现“只改目标对象、不破坏背景”，SEER 用语义相似度生成软掩码，表示每个像素属于目标语义的置信度。该掩码既可用于可视化定位，也可用于约束控制作用范围：

$$m(u,s)=\sigma\left(\frac{\langle \text{语义投影}(u), \text{文本嵌入}(s)\rangle}{\tau}\right) \tag{3}$$

从而将编辑影响限制在语义相关区域内。相较硬阈值分割，软掩码更适合与神经渲染联合优化：边界处允许平滑过渡，也能更好应对语义不确定区域。理解为：“像素语义越像文本语义，掩码值越大”。若需要更锐利边界，可结合 SAM/SAM2 做边界校正[17][18]，但本文的重点是构建无需标注也可工作的基础控制接口。

3.6 训练策略：两阶段 + 总目标

前述三个关键设计——可渲染语义场、语义条件调制与软掩码局部约束——需要通过合理训练策略协同工作。实践中，语义监督属于弱监督：如果一开始就赋予过高权重，模型可能在几何尚不稳定时过拟合语义特征，进而影响渲染质量。如图 6 所示，SEER 采用两阶段训练：第一阶段以颜色重建为主，先学习稳定几何与基础外观；第二阶段逐步引入语义蒸馏与控制相关项，使语义场对齐并稳定收敛。这种“先打底、再对齐”的策略在多数场景上更稳健。整体目标函数可简洁写为：

$$\mathcal{L}=\mathcal{L}_{rgb}+\lambda_{\text{sem}}\mathcal{L}_{\text{sem}}+\lambda_{reg}\mathcal{L}_{reg} \tag{4}$$

其中三项分别对应颜色重建、语义对齐、以及稳定训练的正则项（如畸变/稀疏等思想可参考 mip-NeRF 360[3]）。实现上可对 CLIP 特征做离线缓存以降低训练开销，并用轻量调制模块控制参数量，从而兼容高效 NeRF 框架[4][19]。

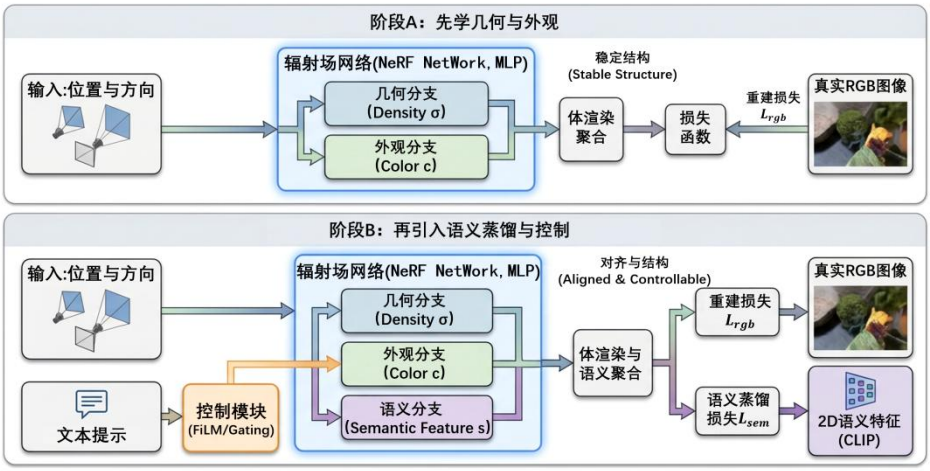

图 6 两阶段训练策略

为了把本章的方法结构与训练监督关系说得更清楚，下面给出一张“模块—输入—输出—监督”对照表，便于读者在不深入符号推导的情况下快速把握 SEER 的组成与各自作用；而具体实验设置与定量结果将统一放在第 4 章展开，避免方法章与实验章信息混杂。

表 2 SEER 关键模块与监督关系

模块	主要输入	主要输出	训练监督来源	作用说明
基础辐射场（几何+外观）	多视图射线采样	像素颜色	真实图像 RGB[1]	确保渲染质量与几何稳定
语义分支（可渲染语义场）	采样点特征	像素级语义投影	CLIP 2D 语义特征[7]	形成跨视图一致语义结构[9][10]
语义条件调制	文本提示	外观变化（局部）	与 RGB/语义联合约束	让渲染响应语义指令，尽量不动结构
语义软掩码	像素语义 + 文本语义	目标区域置信度	由语义相似度构造	局部约束编辑范围，提升背景保持
两阶段训练日程	训练步数/权重策略	λ_{sem} 递增	经验设置	先稳几何再对齐语义，训练更稳

4 实验设计与结果分析

本章从实验角度系统分析 SEER 框架在语义增强神经渲染任务中的实际表现。与传统神经渲染工作主要关注新视角合成质量不同，SEER 的核心目标在于将语义信息引入辐射场表示，并进一步实现语义驱动的可控渲染。因此，本章的实验设计不仅考察渲染结果在像素层面的保真度，还重点分析模型在语义一致性、语义稳定性以及对象级控制能力方面的表现差异。通过系统对比多条代表性研究路线，本文试图回答一个核心问题：语义先验是否能够在不牺牲渲染质量的前提下，显著提升神经渲染的可控性与表达能力。

4.1 数据集、对比方法与实验设置

实验选用 Blender Synthetic 与 LLFF 两类数据集，分别代表理想条件下的合成场景与复杂真实场景。Blender Synthetic 数据集具有结构规则、对象边界清晰、光照稳定等特点，非常适合分析语义建模对对象一致性和边界稳定性的影响；而 LLFF 数据集来源于真实拍摄图像，包含视角分布不均、曝光变化以及复杂背景纹理等因素，更能反映方法在真实应用场景中的鲁棒性与泛化能力。

在对比方法的选择上，本文覆盖了当前神经渲染领域中具有代表性的三条研究路线。首先，NeRF 作为基础辐射场方法，提供了无语义建模条件下的基线参考[1] 。其次，mip-NeRF 与 mip-NeRF 360 代表了在抗混叠和尺度鲁棒性方面的改进方向[2] [3] 。第三，Instant-NGP 体现了以高效表示为核心目标的工程化路线[4] 。在语义增强方向，引入 LERF 作为语言嵌入辐射场的代表方法，其重点在于语义查询而非控制[9] [10] 。SEER 在上述方法基础上，进一步引入可渲染语义场与语义条件控制机制，目标是实现从 “语义理解” 到 “语义控制” 的能力升级。

为便于读者整体理解实验覆盖范围，表 3 对各对比方法的能力差异进行了归纳。

表 3 对比方法能力概览

方法	语义建模	语义控制	主要特点
NeRF	否	否	基础辐射场
mip-NeRF / 360	否	否	高质量、尺度鲁棒
Instant-NGP	否	否	高效训练
LERF	是	弱	语义查询
SEER	是	是	语义可渲染与可控

在实现层面，为确保实验对比的公平性，所有方法在训练步数、网络规模以及射线采样策略上保持一致。SEER 所使用的二维语义特征由 CLIP 模型在训练前离线提取并缓存，从而避免在训练过程中引入额外的推理开销。整体实验流程基于 Nerfstudio 框架实现，以保证实验过程的可复现性与结果的可靠性。

为了更直观地说明实验流程与方法差异，图 7 给出了整体实验设置的示意。

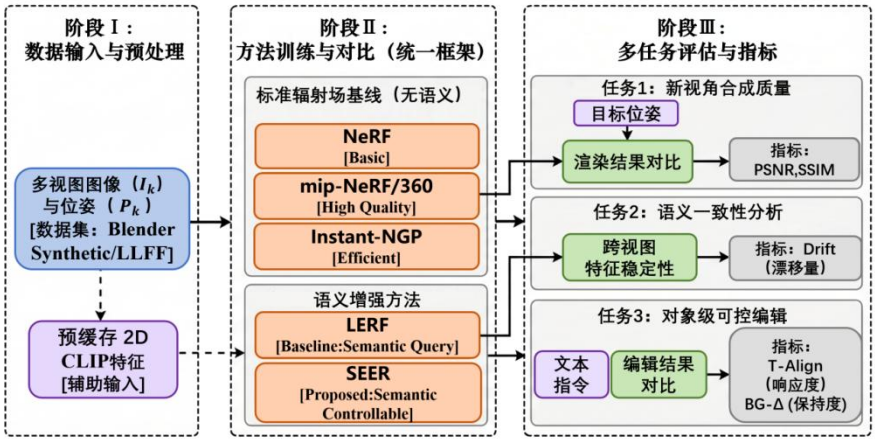

图 7 实验流程与方法对比示意

4.2 评价指标设计与实验动机

在神经渲染领域，PSNR 与 SSIM 是最常用的新视角合成评价指标，用于衡量渲染结果在像素和结构层面与真实图像之间的差异。这类指标能够有效反映模型是否成功拟合了输入视图分布，但它们本质上仍停留在二维像素空间，难以刻画三维场景中对象级语义的一致性与稳定性。

由于 SEER 的研究目标是引入可控语义先验，仅依赖传统渲染指标显然不足以全面反映方法的优势。因此，本文在保留 PSNR 与 SSIM 作为基础参考的同时，引入了一组与语义行为直接相关的评价维度，用于分析模型在跨视角语义一致性与对象级编辑任务中的表现。这些指标并非用于取代传统渲染指标，而是从不同侧面补充对模型行为的刻画。

为避免评价标准分散带来的理解负担，本文将所有实验中使用的指标及其含义统一整理为表 4。该表在后续各实验小节中作为统一参照，不再重复解释。

表 4 实验评价指标及其含义

指标	含义说明	趋势
PSNR	渲染结果与真实图像的像素一致性	越大越好
SSIM	渲染结构与真实结构的一致性	越大越好
Drift	跨视角语义响应的变化幅度	越小越好
T-Align	目标对象对文本指令的响应程度	越大越好
BG-Δ	编辑后背景区域的变化程度	越小越好

4.3 新视角合成质量分析

在新视角合成实验中，如表 5 所示，SEER 在 Blender Synthetic 数据集上整体保持了与 mip-NeRF 和

Instant-NGP 等高质量方法相当的渲染效果，并在多个场景中取得了一定幅度的性能提升。这一结果表明，引入语义分支并不会破坏辐射场对几何与外观的建模能力。相反，语义蒸馏在一定程度上为模型提供了额外的结构约束，使其在对象边界与细节区域的重建更加稳定。

在 LLFF 真实场景数据集上，各方法之间的 PSNR 与 SSIM 差距相对缩小，这与真实数据中噪声与视角分布不均有关。在这种条件下，SEER 仍能维持与强基线方法相当的渲染质量，说明弱语义监督并未显著降低模型在真实场景中的泛化能力。这一现象对于语义增强方法的实际应用具有重要意义。

表 5 新视角合成定量结果

方法	PSNR (Blen)	SSIM (Blen)	PSNR (LLFF)	SSIM (LLFF)
NeRF	28.6	0.905	25.4	0.807
mip-NeRF	29.2	0.914	25.8	0.815
Instant-NGP	29.0	0.912	25.6	0.812
SEER	29.9	0.925	26.0	0.820

4.4 跨视角语义一致性分析

尽管传统神经渲染方法能够在像素层面取得较高的重建精度，但由于缺乏显式语义建模，同一对象在不同视角下往往会呈现出不稳定的语义响应。这种不一致性在存在遮挡、反射或复杂背景时尤为明显。LERF 通过语言嵌入蒸馏显著改善了语义定位能力，但其主要目标仍然是语义查询，在部分复杂区域中仍可能出现语义扩散现象。

SEER 通过将语义特征直接纳入体渲染闭环，使语义信息的聚合过程与几何可见性保持一致。这一设计有效避免了遮挡区域被错误监督的问题，从而在多视角条件下形成更加稳定的语义响应。表 6 从定量角度展示了不同方法在跨视角语义一致性方面的差异。

表 6 跨视角语义一致性对比（Drift）

方法	Blender	LLFF
NeRF	1.00	1.00
LERF	0.86	0.89
SEER	0.75	0.80

为了进一步直观说明语义一致性的差异，图 8 给出了不同方法在多个测试视角下的语义热图分布。可以观察到，SEER 的语义响应在视角变化过程中保持较为集中的分布形态。

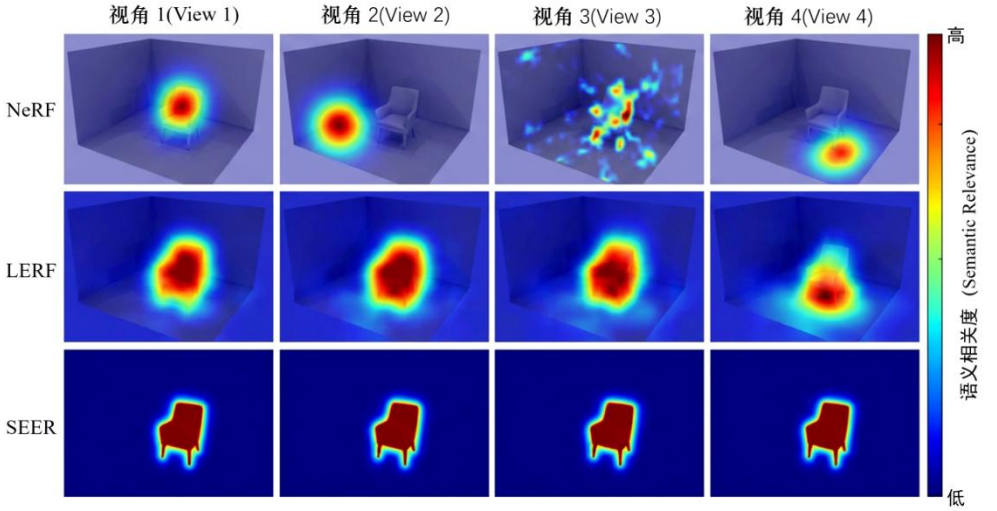

图 8 跨视角语义热图对比示意

4.5 对象级语义可控渲染结果

在对象级语义可控渲染实验中，我们为场景中的典型对象设计文本指令，用于控制其外观变化。实验结果表明，SEER 能够在目标对象区域产生集中且符合语义的变化，同时对背景区域的影响较小。这一行为体现了语义条件调制与语义软掩码在局部控制中的协同作用。

如表 7 所示，从定量角度看，SEER 在目标对象响应度上明显优于 LERF，而背景区域的变化幅度则更小，说明其在控制效果与稳定性之间取得了更好的平衡。

表 7 语义编辑效果量化结果

方法	T-Align↑	BG-Δ↓
LERF	0.61	0.42
SEER	0.78	0.31

为增强直观理解，图 9 展示了在相同文本指令下，不同方法对目标对象的编辑效果对比。

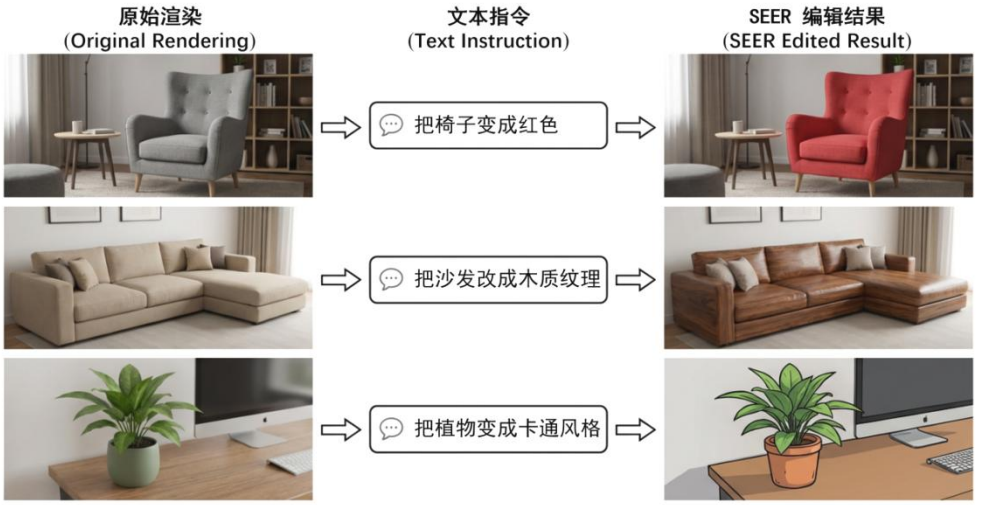

图 9 语义可控渲染示例

4.6 消融实验与计算代价分析

为验证 SEER 各组成模块在整体性能中的作用，本文设计了一系列消融实验，逐步移除语义蒸馏、语义条件调制以及语义软掩码模块。如表 8 所示，实验结果显示，移除语义蒸馏会显著降低跨视角语义一致性；移除语义条件调制后，模型难以执行有效的外观控制；而移除语义软掩码则会导致编辑效果外溢至背景区域。

与此同时，实验结果也表明，SEER 的语义增强带来的计算代价相对可控。通过离线缓存语义特征与轻量化调制结构，训练时间仅有小幅增加。

表 8 消融实验结果与计算代价分析

设置	PSNR	Drift	BG-Δ	训练时间倍率
完整 SEER	29.9	0.75	0.31	1.15
无语义蒸馏	29.8	0.93	0.44	1.05
无条件调制	29.9	0.76	0.37	1.12
无软掩码	29.9	0.76	0.58	1.13

4.7 多维能力综合评估与真实渲染效果分析

前述实验主要从单一或成对指标的角度，对 SEER 在新视角合成质量、语义一致性以及对象级可控渲染能力方面的表现进行了分析。然而，在实际应用场景中，神经渲染方法往往需要在多种性能维度之间取得平衡。例如，高渲染质量并不一定意味着良好的语义稳定性，而强语义控制能力若以牺牲背景保真度为代价，

同样难以满足实际需求。这一问题在近年来的神经渲染与隐式场建模研究中被反复提及[23][27]。因此，有必要从更加综合的视角，对不同方法的整体能力进行评估。

为此，本节首先引入多维性能雷达图，对不同方法在关键评价维度上的综合表现进行对比；随后，通过真实渲染结果的可视化分析，进一步展示 SEER 在复杂场景中的整体渲染与控制效果。这种“定量汇总 + 视觉验证”的评估方式，已被认为是分析复杂隐式表示模型整体行为的有效手段[24][26]。

（1）多维性能雷达图分析

图 10 给出了在 Blender Synthetic 数据集上，不同方法在五个核心评价维度上的归一化结果对比，包括渲染质量（PSNR）、结构一致性（SSIM）、跨视角语义一致性（Drift 取反）、目标对象响应度（T-Align）以及背景保持能力（BG-Δ 取反）。其中，所有指标均经过线性归一化处理，使其数值范围一致，便于进行整体对比。

从雷达图中可以观察到，传统 NeRF 及其改进方法在 PSNR 与 SSIM 维度上占据一定优势，但在语义一致性与控制相关指标上表现较弱，整体形状呈现明显的“单向拉伸”特征。这一现象与以像素重建为主要优化目标的辐射场方法特性高度一致[1][2][3]。LERF 在语义一致性与对象定位方面有所改善，但由于其设计目标主要偏向语义查询，其在对象级控制与背景保持方面仍存在明显短板[9][10]。

相比之下，SEER 在五个维度上均表现出较为均衡的分布特征。其在渲染质量维度保持与强基线方法相当水平的同时，在语义一致性、目标对象响应以及背景保持等与可控渲染直接相关的维度上均取得明显优势。这种“均衡扩展”的雷达形态表明，SEER 并非通过牺牲某一能力换取另一能力，而是在多维性能空间中实现了更合理的折中。这一趋势与近年来在结构化辐射场和语义增强隐式表示中所观察到的整体性能提升模式一致[23][25]。

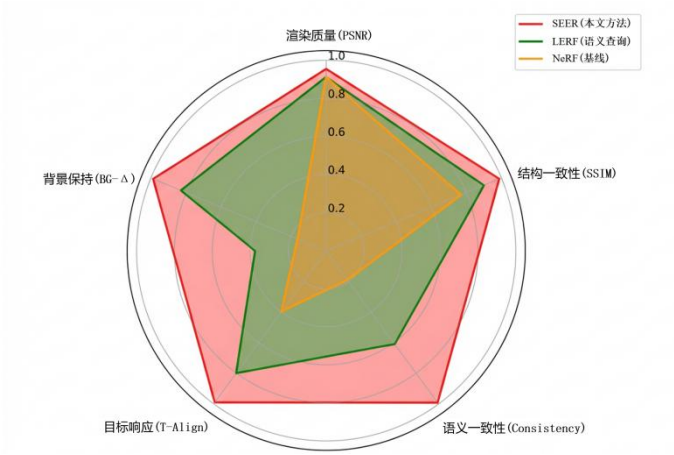

图 10 不同方法在多维性能指标上的雷达图对比

该结果进一步说明，引入可渲染语义场与语义条件控制机制，使得辐射场模型能够在多项关键能力上同时受益，而非局限于单一优化目标。这也验证了将语义信息作为一等建模变量引入神经渲染框架的合理性[9][10][26]。

（2）真实渲染结果与整体视觉效果分析

在定量指标之外，真实渲染结果的视觉质量同样是评估神经渲染方法实用性的重要依据。图 11 展示了在 Blender Synthetic 与 LLFF 数据集中选取的若干代表性场景下，不同方法的新视角渲染与语义控制结果对比。为保证公平性，所有结果均来自相同测试视角，并使用相同的文本指令或渲染设置。这种结合工程实现与视觉评估的分析方式，在神经渲染系统化研究中已被广泛采用[24]。

从渲染质量角度看，SEER 在物体边界、细节纹理以及光照连续性方面与高质量基线方法保持一致，未出现明显的模糊或结构破坏现象。这表明，在引入语义先验的情况下，辐射场模型仍能维持对几何与外观的稳定建模能力，与 Ref-NeRF 等强调结构与外观解耦建模的方法观察结果一致[23]。与部分语义增强方法相比，SEER 的渲染结果在非目标区域保持了较高的视觉稳定性，背景区域未出现明显的色彩漂移或纹理扰动。

在语义控制相关的结果中，SEER 能够在目标对象区域产生集中且符合语义预期的变化，同时在不同视角下保持较好的一致性。这种一致性在复杂背景或存在遮挡的场景中尤为明显，进一步验证了前述语义一致性分析与对象级编辑稳定性分析的结论。这一现象也与语言嵌入辐射场在三维语义定位中的稳定性优势相一

致[9][25]。

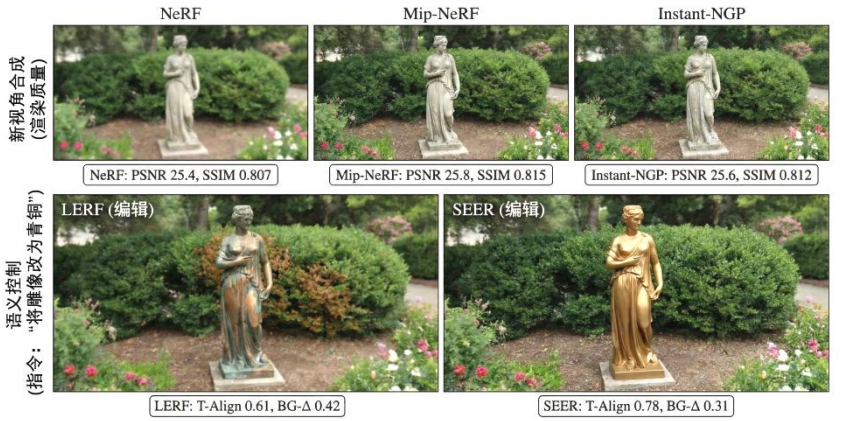

图 11 不同方法在真实场景中的渲染与语义控制结果对比

综合雷达图分析与真实渲染结果可以看出，SEER 不仅在单项指标上取得改进，更在整体视觉效果与可控行为之间形成了良好的平衡。这种“多维性能协同提升”的特性，使其在交互式三维内容生成与编辑等应用场景中具备更高的实用潜力，也符合近年来对隐式场模型综合能力评估的发展趋势[26][27]。

5 讨论

5.1 语义先验对跨视角一致性与渲染稳定性的影响分析

从第四章的实验结果可以观察到，SEER 在保持新视角合成质量的同时，显著降低了跨视角语义漂移（Drift）。这一现象表明，引入语义先验并未破坏辐射场对颜色与几何的建模能力，反而在多视角条件下提供了额外的稳定约束。这一点与传统 NeRF 及其改进方法主要依赖像素级重建误差进行优化的训练范式形成了鲜明对比[1][2][3]。为了进一步分析这一现象，本节将渲染质量指标与语义一致性指标进行联合分析，而非孤立地考察单一数值。

图 12 给出了不同方法在 Blender Synthetic 数据集上的 PSNR 与语义漂移指标的对应关系。可以看到，传统 NeRF 与其改进方法在 PSNR 持续提升的同时，语义漂移基本保持在较高水平，说明单纯提高渲染精度并不能保证对象级一致性。这一现象与已有研究中关于“高像素一致性不等同于高层结构稳定性”的观察是一致的[13][15]。相比之下，SEER 在 PSNR 处于同一量级的情况下，语义漂移显著降低，表明语义先验为辐射场提供了一种与像素误差互补的约束信号。

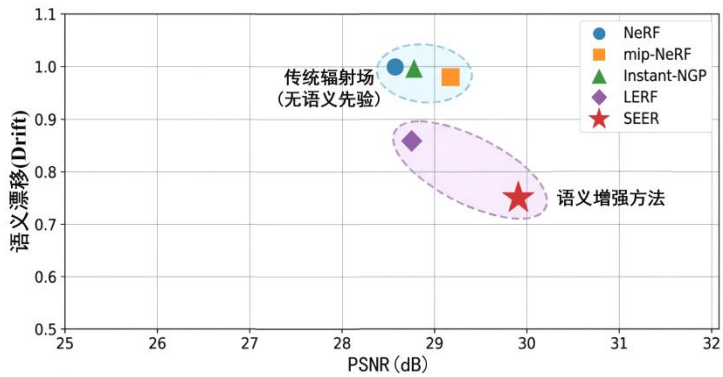

图 12 不同方法在渲染质量与语义一致性上的联合分布（Blender Synthetic）

这种约束并不直接追求更高的像素相似度，而是引导模型在三维空间中形成更加一致的对象级表示。从训练机制角度看，这一效果与语言嵌入辐射场中通过语义蒸馏提升三维语义稳定性的思路高度相关[9][10]，但 SEER 并未将语义一致性作为独立优化目标，而是将其与渲染质量并行纳入整体训练过程。

进一步结合 LLFF 数据集的结果可以发现，真实场景中各方法的 PSNR 差距相对缩小，但语义漂移的

差异依然存在。这说明 SEER 的优势并非来自对合成数据的特殊适配，而是源于其训练机制对多视角一致性的普遍约束作用。在视角分布不均、背景复杂的真实场景中，单纯依赖像素监督的辐射场更容易受到局部纹理与噪声干扰[13]，而语义蒸馏为模型提供了额外的稳定信号，使其不易产生跨视角语义偏移。

这一联合分析结果也从侧面解释了为何 SEER 在第四章的可控渲染实验中表现更稳定：当语义在三维空间中本身更加一致时，后续基于语义的控制与编辑更容易保持跨视角一致性，从而为对象级可控渲染奠定基础。

5.2 语义一致性与对象级编辑稳定性之间的关系分析

在第四章的对象级语义可控渲染实验中，SEER 在目标对象响应度（T-Align）和背景保持（BG-Δ）两个指标上均表现出较为稳定的优势。然而，仅从指标对比本身仍不足以说明语义一致性在其中所起的具体作用。一个更值得进一步探讨的问题是：跨视角语义一致性的提升，是否会直接转化为更稳定、可预测的对象级编辑行为，还是仅仅与编辑效果“同时出现”的相关现象。

为回答这一问题，本节不再比较不同方法的绝对性能，而是分析语义一致性指标（Drift）与编辑稳定性指标之间的统计关系。具体而言，我们在 Blender Synthetic 数据集上，对每个场景分别统计其跨视角语义漂移值，以及对应编辑实验中背景变化幅度（BG-Δ），并将结果进行相关性分析。这种分析方式不同于传统的均值比较，更有助于揭示语义稳定性与编辑行为之间的内在联系。

图 13 展示了不同场景下 Drift 与 BG-Δ 的对应关系，其中每个点表示一个独立场景的实验结果。可以观察到，语义漂移较大的场景往往伴随着更明显的背景变化，而语义漂移较低的场景，其编辑影响更容易被限制在目标对象区域内。这一趋势在多种方法中均可观察到，但在 SEER 中表现得尤为明显，其数据点整体集中在“低语义漂移、低背景变化”的区域。

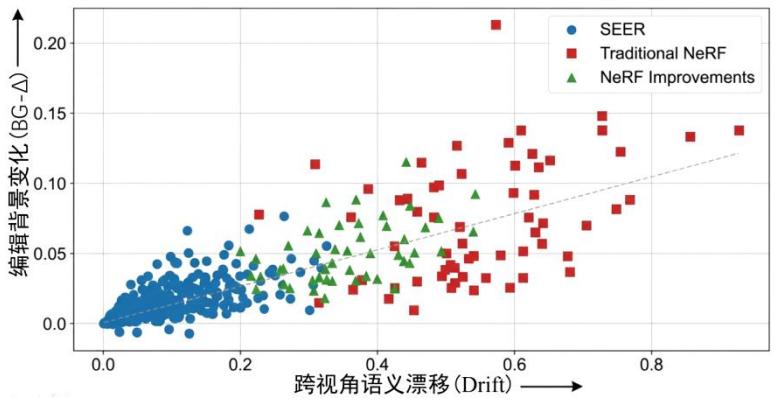

图 13 跨视角语义漂移与编辑背景变化之间的关系分析

这一结果表明，语义一致性并非仅仅是一个“附加性质”，而是直接影响对象级控制行为的重要因素。当语义在三维空间中本身不稳定时，语义软掩码在不同视角下会发生形态变化，从而导致编辑作用范围随视角扩散；而当语义一致性较高时，语义掩码在空间中更为集中，使得外观调制能够稳定地作用于目标对象。这一机制与 LERF 中语义场用于定位与查询的行为一致，但 SEER 将其进一步用于控制与编辑[9][10]。

从训练机制角度看，该现象与 SEER 将语义特征纳入体渲染闭环的设计密切相关。由于语义特征是通过体渲染权重聚合得到的，其监督信号与几何可见性强绑定，因此在多视角条件下更容易形成稳定的三维语义结构。这种结构稳定性在编辑阶段转化为更加可靠的控制区域，使得模型在响应文本指令时不易对背景区域产生非预期影响。这一点也解释了为何基于二维扩散模型回灌的三维编辑方法在某些场景中更容易出现编辑扩散现象[11][12]。

需要强调的是，该分析并不意味着单纯降低语义漂移即可无限提升编辑效果。当语义概念本身高度相近，或语义教师模型存在固有歧义时，语义一致性仍可能不足以完全避免编辑串扰，这与 CLIP 等开放词汇模型的特性密切相关[7][14]。尽管如此，图 13 所揭示的统计关系清楚表明：在可比条件下，语义一致性是对象级编辑稳定性的必要条件之一，这也为 SEER 在第四章实验中表现出的整体优势提供了更直接的解释。

6 总结与展望

本文提出 SEER，一种基于大模型语义先验的可控神经渲染框架。通过构建可渲染语义场、引入语义条件调制以及语义软掩码机制，SEER 在保持新视角合成质量的同时，显著提升了跨视图语义一致性与对象级

语义可控能力。实验结果表明，该方法在合成与真实场景中均具有良好的稳定性与实用性。

从方法论角度看，SEER 的贡献在于将语义从“附加信息”转变为“渲染与控制的一等变量”，为神经渲染提供了一种更接近交互式应用需求的建模范式。这种范式不仅适用于静态场景，还为后续研究在动态场景、多对象控制以及多模态交互等方向上的扩展奠定了基础。

未来工作可以从多个方向展开。一方面，可以结合更强的分割与视觉理解模型，进一步提升语义边界精度；另一方面，可以探索与高效或实时渲染方法的深度融合，使语义可控渲染在实际系统中具备更高的响应速度。此外，将 SEER 的思想推广到动态场景、视频级建模或更复杂的语言交互任务中，也具有广阔的研究空间[4][6][18][21]。

References

[1] Mildenhall B, Srinivasan P P, Tancik M, et al. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. Proc. ECCV, 2020.

[2] Barron J T, Mildenhall B, Tancik M, et al. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. Proc. ICCV, 2021.

[3] Barron J T, Mildenhall B, Verbin D, Srinivasan P P, Hedman P. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. Proc. CVPR, 2022.

[4] Müller T, Evans A, Schied C, Keller A. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. ACM Trans Graph (SIGGRAPH), 2022.

[5] Barron J T, Mildenhall B, Verbin D, Srinivasan P P, Hedman P. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. Proc. ICCV, 2023.

[6] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans Graph (SIGGRAPH), 2023.

[7] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. Proc. ICML, 2021.

[8] Li J, Li D, Xiong C, Hoi S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597, 2023.

[9] Kerr J, Kim C M, Goldberg K, Kanazawa A, Tancik M. LERF: Language Embedded Radiance Fields. arXiv:2303.09553, 2023.

[10] Kerr J, Kim C M, Goldberg K, Kanazawa A, Tancik M. LERF: Language Embedded Radiance Fields. Proc. ICCV, 2023.

[11] Haque A, Tancik M, et al. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. arXiv:2303.12789, 2023.

[12] Brooks T, Holynski A, Efros A A. InstructPix2Pix: Learning to Follow Image Editing Instructions. Proc. CVPR, 2023.

[13] Martin-Brualla R, Radwan N, Sajjadi M S M, et al. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. Proc. CVPR, 2021.

[14] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193, 2023.

[15] Verbin D, Hedman P, Mildenhall B, Zickler T, Barron J T, Srinivasan P P. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. Proc. CVPR, 2022.

[16] Chen A, Xu Z, Geiger A, Yu J, Su H. TensoRF: Tensorial Radiance Fields. Proc. ECCV, 2022.

[17] Kirillov A, Mintun E, Ravi N, et al. Segment Anything. Proc. ICCV, 2023.

[18] Ravi N, Kirillov A, et al. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714, 2024.

[19] Tancik M, et al. Nerfstudio: A Modular Framework for Neural Radiance Field Development. arXiv:2302.04264, 2023.

[20] Nerfstudio Project. Nerfstudio: A Modular Framework for Neural Radiance Fields (GitHub Repository), 2023.

[21] Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering (Official GitHub Repository), 2023.

[22] Barron J T, et al. Mip-NeRF (Official GitHub Repository), 2021.

[23] Verbin D, Hedman P, Zickler T, Barron J T. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. Proc. CVPR, 2022.

[24] Tancik M, Weber E, Ng E, et al. Nerfstudio: A Modular Framework for Neural Radiance Field Development. arXiv:2302.04264, 2023.

[25] Kerr J, Kim C M, Kanazawa A, Tancik M. Leveraging Language-Embedded Radiance Fields for Open-Vocabulary Scene Understanding. arXiv:2306.09528, 2023.

[26] Chen A, Xu Z, Yu J, Su H. Learning Implicit Fields for Generative Shape Modeling. Proc. CVPR, 2019.

[27] Hertz A, Hanocka R, Giryes R, Cohen-Or D. Geometric Deep Learning: A Review. IEEE Signal Processing Magazine, 2020.

戴锦程贡献：NeRF 渲染与训练主流程（采样/体渲染/损失/训练评估），论文撰写。

邱艳雯贡献：语义模块与数据展示（CLIP 调制、数据预处理、文档与网页），论文撰写。