测绘学报　2026 年　第 55 卷　第 5 期：1001—1595
Acta Geodaetica et Cartographica Sinica　xb.chinasmp.com
中国测绘学会 Chinese Society for Geodesy Photogrammetry and Cartography
测绘出版社 SURVEYING AND MAPPING PRESS

时空信息显式引导的高分光学遥感影像可控生成方法

时天东[1]，赵　玲[1]，赵文豪[2]，齐　霁[3]，崔　浩[4]，彭程里[1]，张新长[3]

1. 中南大学地球科学与信息物理学院，湖南 长沙 410083；
2. 国家基础地理信息中心，北京 100830；
3. 广州大学地理科学与遥感学院，广东 广州 510006；
4. 武汉大学测绘遥感信息工程全国重点实验室，湖北 武汉 430079

收稿日期：2025-12-19；修回日期：2026-04-19　中图分类号：P237　文献标识码：A
文章编号：1001-1595(2026)05-0894-15
基金项目：国家自然科学基金(42571533；42371406)；空间基准全国重点实验室开放基金(SKLSD2026-KF-27)
第一作者简介：时天东(1996—)，男，博士生，研究方向为遥感影像智能解译与可控生成。　E-mail：csushitd@csu.edu.cn
通信作者：齐霁　E-mail：jameschi95@foxmail.com

摘　要：遥感地物的视觉表现受季节演变与地域差异影响显著，如何增强生成模型的时空可控性，精准复现特定时空背景下的地物特征，是当前高分光学遥感影像生成领域面临的关键挑战。现有研究在多源时空信息的编码方式，以及时空特征与视觉特征交互引导方式这两个方面存在局限性，难以准确建模时空条件与地物视觉形态的精确映射关系。对此，本文提出了一种面向时空可控的高分光学遥感影像生成方法。首先，设计了顾及属性差异的多源时空信息编码方式，利用异构频率编码与独立投影将不同属性的时空信息转化为解耦的高维特征表示，以建模多源时空信息的独特属性。其次，设计基于解耦注意力的时空-文本联合交互引导机制，采用独立并行注意力分支促进时空特征与视觉特征的交互引导，在不干扰文本引导生成的同时，充分发挥时空信息在影像生成过程中的约束作用。此外，利用低秩适配训练策略，仅微调少量参数即实现了领域知识的高效迁移与基础生成能力的完整保留。在覆盖我国 7 个典型区域的大规模数据集上的试验表明，本文方法在时空分布一致性和结构纹理一致性上较现有先进方法分别提升了 46.69%和 14.67%，证实了该框架在多样时空场景下的生成可控性与泛化潜力。

关键词：时空智能；遥感影像；生成模型；扩散模型；深度学习；时空信息

遥感影像为城市规划[1]、灾害监测[2]、环境保护[3]等任务提供了信息支撑。此外，作为时空信息的主要载体，遥感影像已成为发展时空智能研究的关键要素[4-6]。然而，受限于卫星重访周期和地物固有的分布规律，人工构建的遥感影像样本库不可避免地存在分布失衡现象[7-8]。这种分布失衡导致数据驱动的智能解译模型难以在多样化时空场景下取得稳定的性能。对此，利用生成模型按需合成符合特定时空分布的遥感影像，是低成本构建均衡样本库并提升解译模型泛化能力的关键途径；同时也能为变化检测任务和数字孪生环境构建提供丰富的样本支撑与虚拟环境支持[9-11]。

由于高分光学遥感影像具有高空间分辨率、广阔地理覆盖度及复杂地物布局等特点，早期采用一步到位建模方式的生成模型[12-13]难以精准捕捉其复杂的数据分布模式，导致生成的影像细节模糊且缺乏灵活的条件引导生成机制，实用价值受限。近年来，扩散模型[14]凭借多步迭代的建模方式，在生成质量上取得了显著突破。然而，缺乏条件约束的随机生成无法满足实际任务对特定场景影像的

引文格式：时天东，赵玲，赵文豪，等. 时空信息显式引导的高分光学遥感影像可控生成方法[J]. 测绘学报，2026，55(5)：894-908. DOI：10.11947/j.AGCS.2026.20250529.
SHI Tiandong，ZHAO Ling，ZHAO Wenhao，et al. Controllable generation of high-resolution optical remote sensing image explicitly guided by spatio-temporal information[J]. Acta Geodaetica et Cartographica Sinica，2026，55(5)：894-908. DOI：10.11947/j.AGCS.2026.20250529.

需求。得益于扩散模型强大的条件引导生成能力，其能够将外部引导信息编码以实现对生成内容的准确控制[15]。现有的高分光学遥感影像可控生成研究主要分为两类。第一类工作利用表示几何边界的图像来增强对地物空间布局的控制。其中，针对语义分割任务的数据稀缺问题，一些研究利用语义掩码作为条件，并结合特征提取与交互模块，实现了像素级布局可控的影像生成[16-19]。针对目标检测任务的数据稀缺问题，AeroGen[20] 与 CC-Diff＋＋[21] 利用目标检测框图作为条件，实现了区域级布局可控的影像生成。第二类工作利用文本描述来增强对地物类别的控制。文献[22]结合大语言模型和扩散模型实现了逼真的灾后影像合成。RSDiff[23] 采用两阶段的级联生成架构实现了高分辨率的光学遥感影像生成。进一步，Text2Earth[24] 和 MetaEarth[25] 构建了全球范围的影像与文本配对数据集，并结合自级联生成框架，显著提升了对全球范围内多样化场景的影像生成效果。此外，DiffusionSat[26] 将经纬度坐标、像元分辨率、采样时间等条件作为引导信息，探索了时空背景对光学遥感影像生成的引导作用。

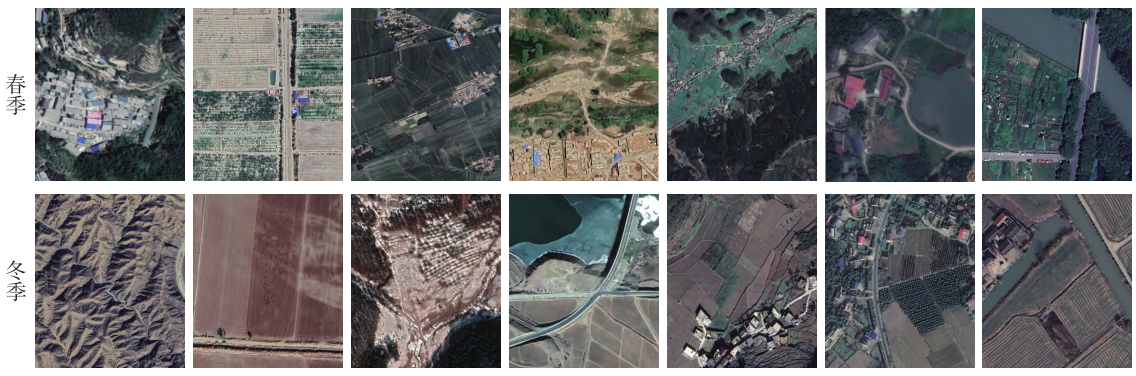
尽管上述研究在地物类别与布局控制上取得了显著进展，但忽略了高分光学遥感影像所固有的时空异质性，导致生成结果常出现物候错乱或地域风格失真。究其原因，现有方法存在两方面局限。首先是时空信息编码方式单一。现有方法多采用统一的多层感知机投影（multilayer perceptron，MLP）或频率编码，难以兼顾多源时空信息的差异化属性，导致时空特征的分布规律被破坏。其次是多模态特征间的交互引导存在干扰。现有方法采取特征拼接或相加的方式实现时空特征与视觉特征的交互引导，这两种交互引导方式都存在特征干扰问题，导致模型难以在保持地物语义内容的同时，实现时空特征对地物纹理和风格的准确引导。

针对上述局限，本文提出一种面向时空可控的高分光学遥感影像生成方法。首先，设计顾及属性差异的异构频率编码策略，将多源时空信息转化为解耦的高维特征表示，以准确建模其独特属性。其次，构建基于解耦注意力的时空-文本联合交互引导机制，通过创建独立并行的时空引导分支，在不干扰文本语义引导的前提下，实现时空特征对地物视觉纹理和风格的准确引导。试验结果表明，本文方法在时空分布一致性和结构纹理一致性上较现有先进方法分别提升了 46.69％和 14.67％，证实了本文方法在提升高分光学遥感影像时空可控生成方面的显著优势。

1　研究区域与数据集构建

1.1　影像与时空信息采集

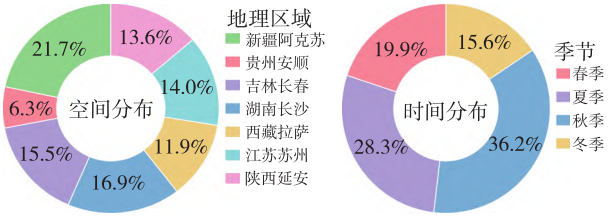
为构建具有时空多样性的高分光学遥感影像数据集，本文首先选取了我国 7 个典型区域（图 1）。这些区域覆盖了不同的气候带与地貌特征，且影像采集时间覆盖全部 12 个月份，以确保影像在时间、空间、地貌和气候 4 个维度上的多样性。本文采用层次化采样策略，将每个区域进一步划分为城市、工业、乡村和自然 4 个子区域。这种分层采样策略旨在捕获同一地理背景下，由人类活动强度不同所导致的地表覆盖差异。最终从数十万张影像中筛选得到 57 920 张 1024×1024 像素的高分辨率影像，影像的空间分辨率涵盖 0.3～2.1 m 范围。此外，本文同步收集了每张影像的采样月份和中心点经纬度坐标，并根据影像所属的地理区域，为每张影像分配了 1 个气候类型标识，构成本文方法所使用的时空信息。影像的空间分布和时间分布统计如图 2 所示。



(a) 陕西延安　(b) 新疆阿克苏　(c) 吉林长春　(d) 西藏拉萨　(e) 贵州安顺　(f) 湖南长沙　(g) 江苏苏州

图 1　具有时空多样性的影像采集区域

Fig. 1　Image sampling regions with temporal and spatial diversity



(a) 空间分布 (b) 时间分布

图 2 影像采样空间和时间的统计

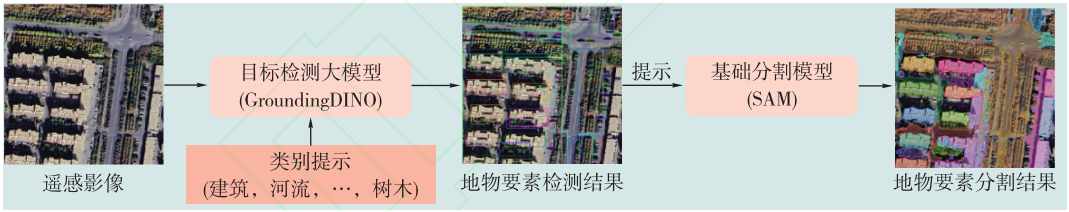
Fig. 2 Statistics of image sampling regions and times

1.2 详细文本描述构建

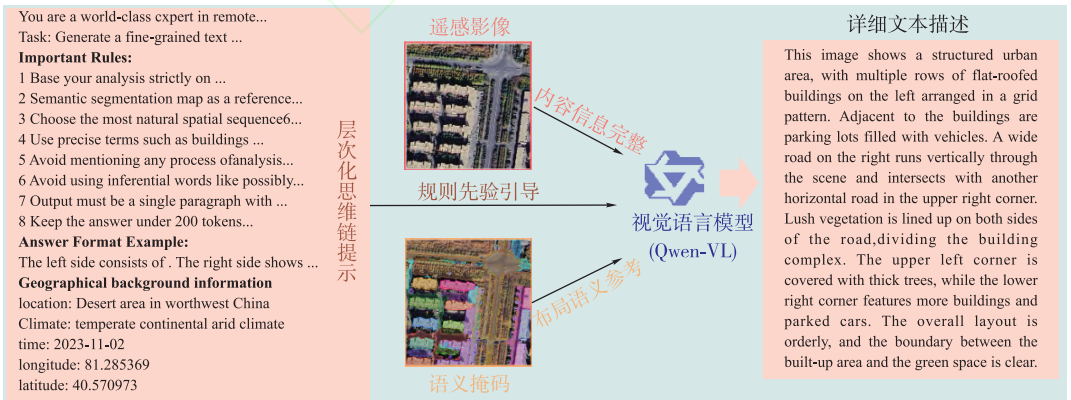
本文方法的训练过程依赖大规模且高质量的高分光学遥感影像与文本配对数据集。由于本文采集的影像规模大，且影像中通常包含丰富的地物要素与复杂的空间布局，采用人工标注来获取文本描述的方式难以满足生成模型训练过程对大量数据的需求。因此，需要设计自动化方法来高效构建相应文本描述。然而，直接调用视觉语言模型[27]为复杂的高分光学遥感影像生成文本描述，容易产生地物内容遗漏或地物布局错乱的错误回答。为此，为高效且准确地构建详细文本描述用于训练生成模型，本文提出一种结合多源输入与层次化思维链提示的详细文本描述生成方法(图 3)。该方法将原始影像、地物要素分割结果及层次化思维链提示共同输入视觉语言模型，协同引导其生成准确的详细描述文本。总体流程分为两个阶段。

第一阶段是视觉基础模型驱动的地物要素细粒度感知结果生成(图 3(a))，采用检测和分割的级联策略获取地物语义掩码。首先，利用目标检测大模型[28]，根据预设类别对影像中的地物要素进行初步定位。然后，将检测结果作为提示输入基础分割模型[29]，获得像素级的语义掩码。该掩码将作为后续文本生成的空间布局先验。

第二阶段是多源输入和思维链提示引导下的详细文本描述生成(图 3(b))，将原始遥感影像、地物要素分割结果和思维链提示指令共同输入视觉语言模型。其中，地物要素分割结果提供粗粒度的空间布局参考，原始影像负责提供完整且真实的视觉信息。此外，本文设计了层次化的思维链提示来引导视觉语言模型进行推理。在提示中明确加入了“作为参考的分割掩码图可能是错误的，最终输出必须包含遥感影像中实际可见的内容”这一显式约束，强制模型在推理时以原始影像为基准。这种层次化的推理过程有效降低了地物要素分割结果不准确带来的干扰，最大限度确保最终生成文本描述的准确性。



(a) 视觉基础模型驱动的地物要素细粒度感知



(b) 多源输入和思维链提示引导下的详细文本描述生成

图 3 基于视觉语言模型的详细文本生成

Fig. 3 Fine-grained text generation with visual language model

2 研究方法

2.1 面向时空可控的高分光学遥感影像生成总体框架

针对现有方法在多源时空信息编码中造成的属性失真及多模态特征交互过程中的干扰等问题，本文提出一种面向时空可控的高分光学遥感影像生成方法(图 4)。综合考虑生成可控性与实现可行性，本文将时空信息聚焦于对地物视觉呈现起决定

性作用的 3 组核心要素：表征物候周期的月份、表征空间分布的经纬度坐标，以及表征宏观地理先验的气候类型标识。针对离散的气候类型描述，本文预先构建映射表将其转化为数值索引，从而确保输入信息在数据格式上的统一性。在此基础上，通过多源时空信息编码以及时空-文本联合引导生成两个关键阶段实现高质量的可控生成，具体流程如下。

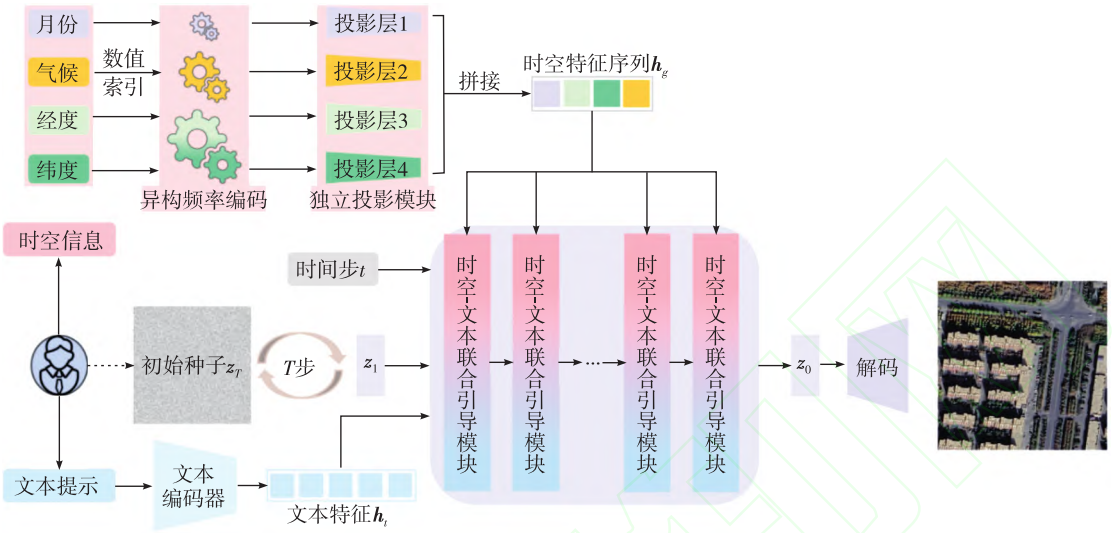


图 4　面向时空可控的遥感影像生成框架

Fig. 4　Temporal and spatial controllable remote sensing image generation framework

（1）多源时空信息编码阶段旨在构建准确且解耦的时空特征表示。为准确建模多源时空信息的差异化属性，本文设计了基于差异化维度的异构频率编码策略，并引入独立投影模块，将不同属性的时空信息映射至独立的语义子空间，最终转化为时空特征序列作为生成过程的显式约束条件。

（2）时空-文本联合引导生成阶段旨在实现解耦的多模态特征交互。在去噪网络内部原有的文本交互引导分支基础上，构建独立并行的时空交互引导分支。两者共同构成了基于解耦注意力的时空-文本联合引导模块，这种解耦的特征交互机制使得时空特征与文本特征在各自独立的特征空间内分别与视觉特征交互，实现了时空特征与文本特征的协同交互引导，避免了互相干扰。

2.2　多源时空信息异构频率编码策略

为使扩散模型准确理解多源时空信息的属性含义，建立有效的多源时空信息编码机制至关重要。为此，本文提出一种基于差异化维度的异构频率编码策略。首先，对多源时空信息的原始数值进行归一化并引入尺度因子进行线性缩放，以消除多源时空信息的量纲差异并增强频率编码对数值变化的感知能力。其次，针对多源时空信息的内在属性差异，采取差异化维度的异构频率编码，准确建模不同属性时空信息的内在变化规律

$$\begin{aligned}\boldsymbol{f}_{\omega}(m_k)=&(\cos(m_k\cdot\omega^{-1/d_k}),\sin(m_k\cdot\omega^{-1/d_k}),\\&\cos(m_k\cdot\omega^{-2/d_k}),\sin(m_k\cdot\omega^{-2/d_k}),\cdots,\\&\cos(m_k\cdot\omega^{-n/d_k}),\sin(m_k\cdot\omega^{-n/d_k}))\end{aligned}\tag{1}$$

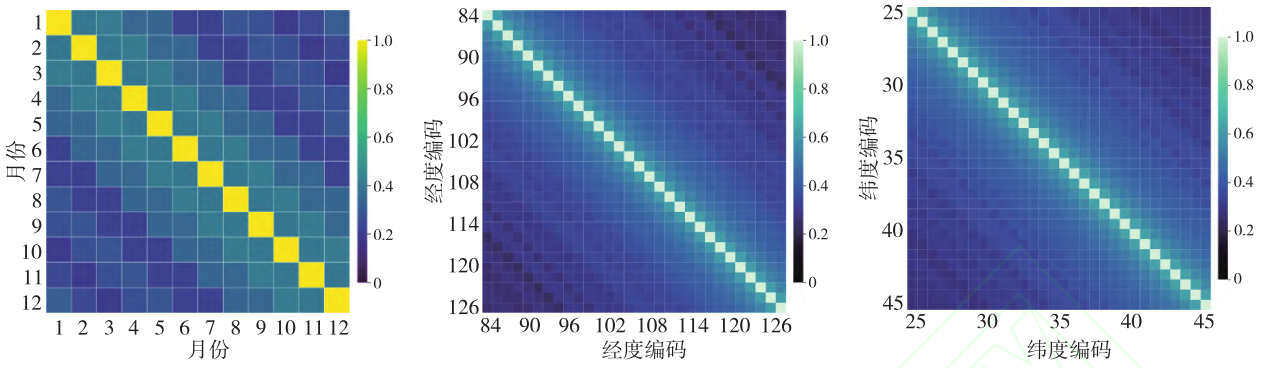
式中，m_k 表示不同类型的时空信息；d_k 是其对应的编码维度；ω 是常数。这种编码策略通过组合不同频率的正余弦函数分量，构建了丰富的多尺度特征表示，能够有效满足本文对于多源时空信息的编码需求。

具体而言，月份信息的核心特征是周期性，即 12 月与 1 月的物候特征具有语义相似性。为此，本文为其分配较小的编码维度，旨在降低高频细节丰富度，使频率编码聚焦于反映数据整体结构的低频特征。如图 5(a)所示，相邻月份及首尾月份的频率编码间维持了较高的余弦相似度，有效捕捉了周期性先验。相对地，经纬度坐标的核心特征是平滑连续性。为此，本文为其分配较大的编码维度，旨在引入更丰富的高频分量，使得频率编码对经纬度坐标的变化具备较高的敏感性。如图 5(b)和图 5(c)所示，经纬度坐标频率编码的余弦相似度随距离增加而平滑衰减，实现了对经纬度坐标平滑连续性的准确表示。

进一步，为统一异构频率编码维度并解耦不同时空信息的语义，本文为每种时空信息分别创建独立的轻量化投影模块 $\boldsymbol{P}_k$，将每类时空信息的频率编码映射到独立的语义子空间，如式(2)所示

$$\boldsymbol{e}_k = \boldsymbol{P}_k(\boldsymbol{f}_\omega(m_k)) \tag{2}$$

随后将它们拼接为时空特征序列 $\boldsymbol{h}_g = (\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{e}_4)$，为后续的引导生成提供准确且解耦的时空特征表示。



(a) 月份编码余弦相似性　(b) 经度编码余弦相似性　(c) 纬度编码余弦相似性

图 5　月份编码和坐标编码的余弦相似性

Fig. 5　Cosine similarity between month coding and coordinate coding

2.3　基于解耦注意力的时空-文本联合交互引导机制

获得时空特征序列 $\boldsymbol{h}_g$ 后，如何将其有效融入扩散模型以实现对影像纹理和风格的准确引导，是时空特征交互引导的核心挑战。从成像机理来看，时空信息主要影响地物的视觉纹理和风格，并非影响地物的语义内容。为此，本文基于扩散变换器架构[30]（diffusion transformer，DiT），提出一种基于解耦注意力的时空-文本联合交互引导机制，如图 6 所示。在保留扩散模型去噪网络内部原有文本交互引导分支的同时，构建了独立并行的时空交互引导分支，二者共同构成时空-文本联合引导模块。这种解耦并行的设计避免了基于特征拼接或相加的交互引导方式容易引发的多模态特征干扰问题，确保时空特征具有独立的特征交互空间，从而在保留扩散模型原有文本引导生成能力的前提下，实现时空特征与视觉特征的深度交互。

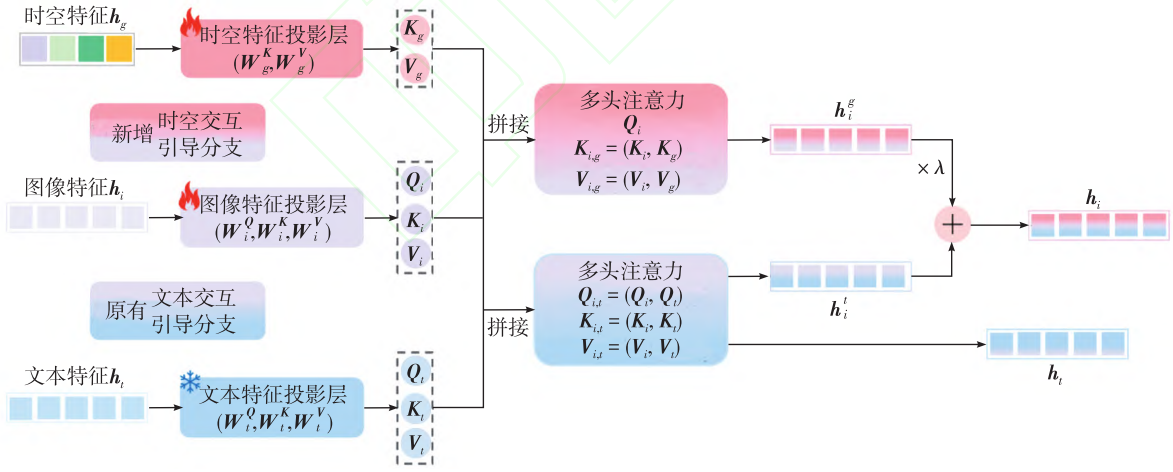


图 6　基于解耦注意力的时空-文本联合引导模块

Fig. 6　Spatio-temporal-text joint guidance module based on decoupled attention

其中，文本交互引导分支是文生图扩散模型内部原有组件，通过注意力机制建模文本特征和图像特征之间的交互。图像特征 $\boldsymbol{h}_i$ 和文本特征 $\boldsymbol{h}_t$ 分别通过各自的投影层进行投影与拼接，形成查询向量序列 $\boldsymbol{Q}_{i,t} = (\boldsymbol{Q}_i, \boldsymbol{Q}_t)$、键向量序列 $\boldsymbol{K}_{z,t} = (\boldsymbol{K}_i, \boldsymbol{K}_t)$ 和值向量序列 $\boldsymbol{V}_{i,t} = (\boldsymbol{V}_i, \boldsymbol{V}_t)$。随后通过多头注意力计算并将输出结果重新拆分，得到融合了文本语义的更新后图像特征 $\boldsymbol{h}_i^t$ 和相应的更新后文本特征 $\boldsymbol{h}_t$

$$(\boldsymbol{h}_i^t, \boldsymbol{h}_t) = \mathrm{Attention}(\boldsymbol{Q}_{i,t}, \boldsymbol{K}_{i,t}, \boldsymbol{V}_{i,t}) = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{i,t} \cdot \boldsymbol{K}_{i,t}^{\mathrm{T}}}{\sqrt{d}}\right) \cdot \boldsymbol{V}_{i,t} \tag{3}$$

时空交互引导分支为新增核心组件，通过注意力机制建模时空特征和图像特征之间的交互，从而实现时空特征对生成影像中地物纹理和风格的引导。该分支复用图像投影 $\boldsymbol{Q}_i$ 作为查询向量，并引入两个独立的可训练投影矩阵 $\boldsymbol{W}_g^K$ 和 $\boldsymbol{W}_g^V$ 将时空特征

序列$\boldsymbol{h}_g$映射为相应的键向量$\boldsymbol{K}_g$和值向量$\boldsymbol{V}_g$。为使查询向量$\boldsymbol{Q}_i$能够协同利用视觉特征与外部的时空特征，在保持影像内容一致性的同时自适应地聚合时空特征信息，本文将两者的键向量和值向量进行拼接，形成键向量序列$\boldsymbol{K}_{i,g}=(\boldsymbol{K}_i,\boldsymbol{K}_g)$和值向量序列$\boldsymbol{V}_{i,g}=(\boldsymbol{V}_i,\boldsymbol{V}_g)$。随后经由多头注意力计算，图像特征自适应地聚合相关的时空特征，得到融合了时空约束的更新后图像特征$\boldsymbol{h}_i^g$

$$\boldsymbol{h}_i^g=\mathrm{Attention}(\boldsymbol{Q}_i,\boldsymbol{K}_{i,g},\boldsymbol{V}_{i,g})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i\cdot\boldsymbol{K}_{i,g}^{\mathrm{T}}}{\sqrt{d}}\right)\cdot\boldsymbol{V}_{i,g}\tag{4}$$

最后，通过缩放因子λ对这两个分支的输出进行加权融合，获得最终更新后的图像特征$\boldsymbol{h}_i$。这种解耦并行的交互引导机制，既实现了多源时空信息对地物纹理和风格的约束，又不干扰扩散模型原有的文本语义引导生成能力

$$\boldsymbol{h}_i=\boldsymbol{h}_i^t+\lambda\cdot\boldsymbol{h}_i^g\tag{5}$$

2.4　基于低秩适配的训练策略

为将预训练的文生图扩散模型有效迁移到高分光学遥感影像领域，本文采用低秩适配微调[31]方式进行训练。首先，冻结预训练模型内部的文本特征投影层，保留其原有的文本引导能力。其次，在图像特征投影层中添加可训练的低秩适配矩阵，用于更新图像特征投影层的权重以学习高分光学遥感影像特有的视觉特征。最后，对新增的时空信息独立投影模块$\boldsymbol{P}_k$和时空交互引导分支中的投影矩阵$\boldsymbol{W}_g^k$与$\boldsymbol{W}_g^v$进行参数更新，以建立时空信息与地物视觉特征之间的映射关系。在训练阶段，使用标准的流匹配优化目标[30]进行端到端训练。

$$L=E_{\boldsymbol{z}_0,t,\boldsymbol{\epsilon}}\left[\left\|(\boldsymbol{\epsilon}-\boldsymbol{z}_0)-\epsilon_\theta(\boldsymbol{z}_t,t,\boldsymbol{h}_g,\boldsymbol{h}_t)\right\|^2\right]\tag{6}$$

式中，$\boldsymbol{\epsilon}$是从标准高斯分布采样的噪声；ϵ_θ表示去噪网络；t表示时间步；$\boldsymbol{z}_0$表示训练期间对原始图像的编码表示；$\boldsymbol{z}_t$表示对$\boldsymbol{z}_0$添加相应的噪声。

3　试验与分析

3.1　试验设置

本文采用分层随机抽样策略，按 9∶1 的比例将完整数据集划分为训练集与测试集，以保证地理分布与地物类别的同分布性。方法基于 Stable Diffusion3 模型拓展实现，试验在 Ubuntu 18.04 操作系统和两个 A6000 GPU 计算平台上完成。训练批量大小设置为 8，采用 AdamW 优化器，初始学习率设为4×10^{-4}并结合余弦衰减策略迭代 50 000 步。低秩适配矩阵的秩设为 1024。在推理生成阶段，生成步数设置为 30 步，无分类器引导系数设置为 5.5。在时空信息的异构频率编码中，缩放因子设为 1000，月份、气候及经纬度的编码维度分别配置为 128、256 和 512。

为验证本文方法的有效性，选取基于文本提示的时空信息编码策略（记为“文本编码法”）、CRS-Diff[32]、DiffusionSat[26]及 Text2Earth[24]作为对比方法。CRS-Diff 采用统一的 MLP 投影对多源时空信息进行编码，之后将时空特征与文本特征拼接并采用交叉注意力实现与视觉特征的交互引导。DiffusionSat 基于统一的频率编码对多源时空信息进行编码，然后将时空特征与视觉特征直接相加进行交互引导。Text2Earth 接收分辨率作为元信息，且编码方式和特征交互引导方式与 DiffusionSat 相同。为此，本文将时空信息与文本提示拼接作为其输入，以此在输入层面保持对齐。

本文从时空分布、结构纹理和语义内容 3 个维度评估生成影像与输入条件的一致性。时空分布一致性采用 FID（fréchet inception distance）指标[33]，该值越低表明生成影像的特征分布与相同时空背景下真实影像的特征分布更为一致。结构纹理一致性采用多尺度结构相似性（multi-scale structural similarity，MS-SSIM）指标[34]，该值越高表明生成影像的纹理结构越逼近真实影像。语义内容一致性采用 CLIP Score（contrastive language-image pre-training score）指标[35]，该值越高表明生成影像中的地物内容与文本描述的语义内容一致性越高。

3.2　试验结果与分析

各方法在测试集上的定量评估结果见表 1。本文方法在 FID、MS-SSIM 和 CLIP Score 这 3 项指标上均取得最优性能。具体而言，本文方法的 FID 指标降至 3.457 2，较次优的 DiffusionSat 优化了 46.69%；MS-SSIM 指标升至 0.225 8，较次优的 CRS-Diff 提升 14.67%。同时，本文方法生成的影像与文本提示具有更高的语义内容一致性。

表 1　各方法在测试集上的定量评估结果

Tab. 1　Quantitative evaluation results of each method on the test dataset

方法	FID	MS-SSIM	CLIP Score
文本编码法	6.858 0	0.150 4	23.948 6
CRS-Diff	7.027 0	0.196 9	25.595 7
DiffusionSat	6.485 0	0.167 0	25.587 7
Text2Earth	13.523 4	0.189 9	24.787 2
本文方法	3.457 2	0.225 8	25.770 5

由表 1 可知，CRS-Diff 与 DiffusionSat 在时空分布一致性与结构纹理一致性上均显著弱于本文方法，反映了它们在时空信息编码与多模态特征交互引导机制两方面的局限。而文本编码法在纹理结构一致性和语义内容一致性上均表现最差，并且时空分布一致性也显著弱于本文方法。这证实了简单的文本化处理无法建立时空信息与地物视觉特征之间的映射关系。此外，Text2Earth 因缺乏显式的时空信息输入和交互引导机制，其 FID 与 MS-SSIM 显著弱于本文方法。这证实了对于时空可控的高分光学遥感影像生成任务，设计时空信息编码机制与多模态特征交互引导机制比单纯扩大数据规模更为关键。

图 7 展示了本文方法与对比方法的生成结果。整体而言，本文方法在语义内容、结构纹理及时空分布一致性上均优于对比方法。以第 5 列的“6 月温带大陆性季风气候，农田场景”为例，CRS-Diff 与 DiffusionSat 生成的影像未能准确反映该时空背景的物候特征；Text2Earth 则随机生成了不符合该时空背景的影像。相反，本文方法准确呈现了夏季农作物的视觉特征，且生成的矩形田块纹理结构清晰，符合该时空背景下的地物视觉特征。这些定量和定性结果表明，本文提出的多源时空信息异构频率编码策略与基于解耦注意力的时空-文本联合交互引导机制，能够准确捕捉并建模遥感影像在不同时空背景下的纹理细节与风格特征，提升生成影像的真实感。同时，完整保留了模型原有的文本语义引导能力。

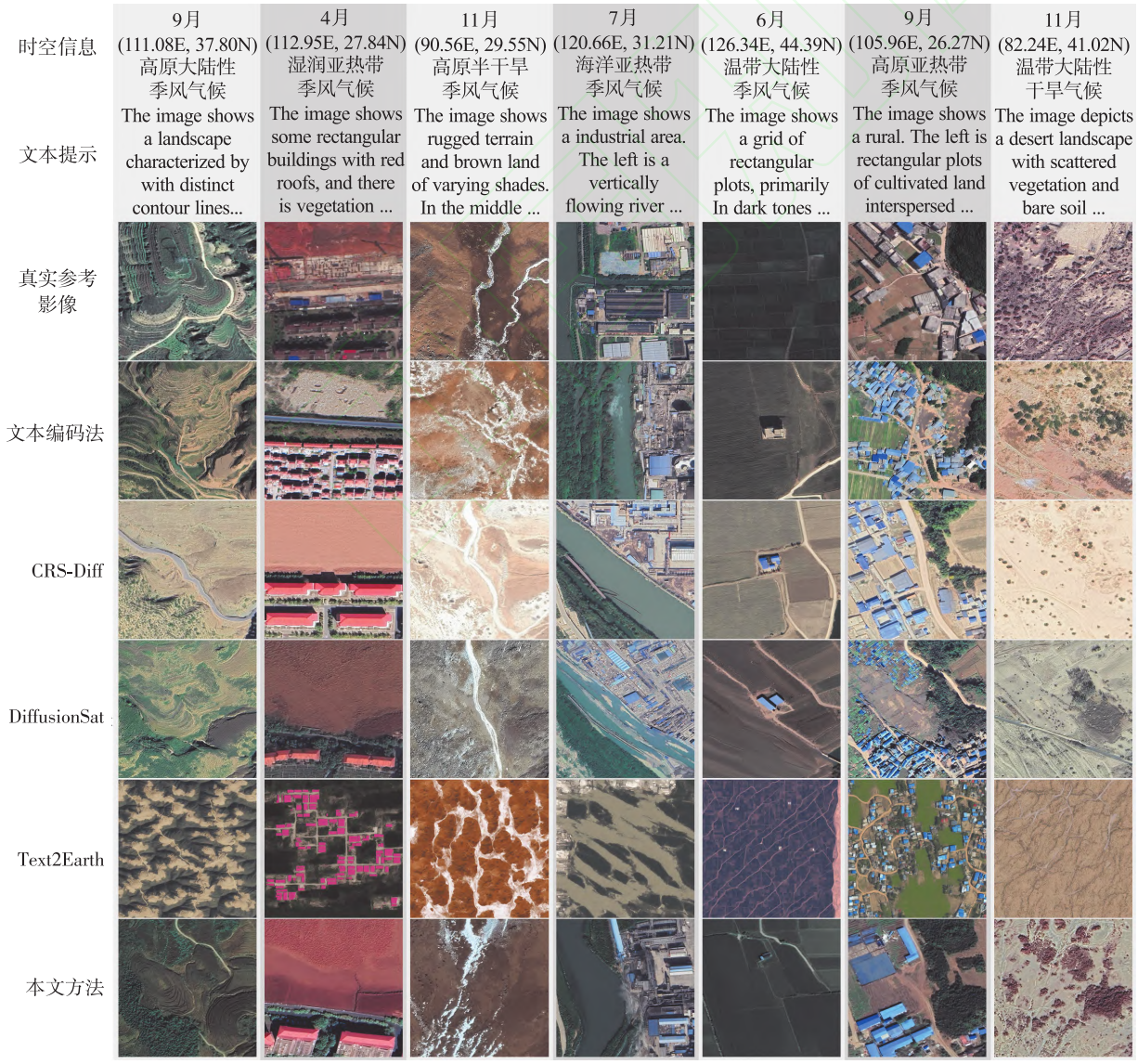
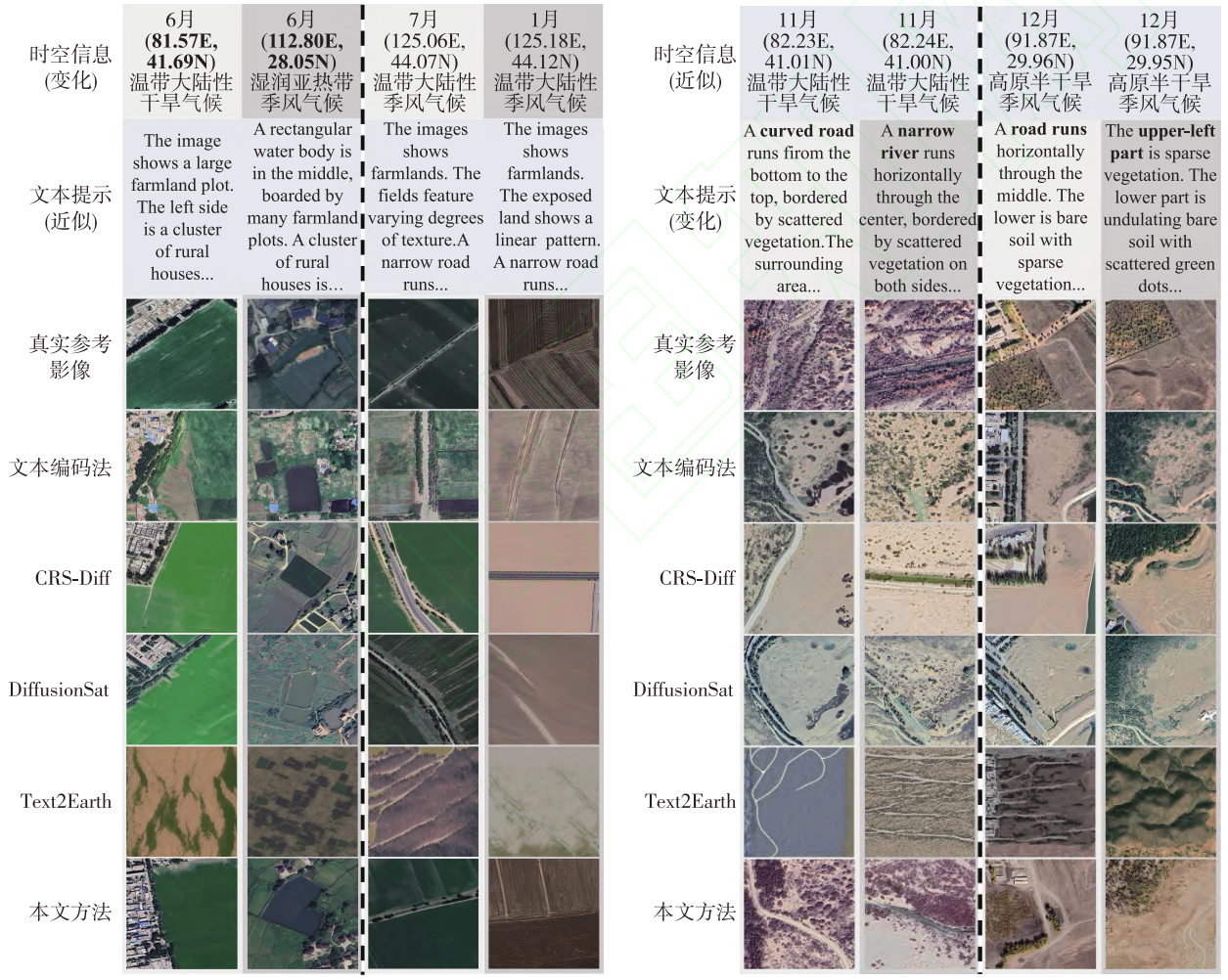


图 7　使用不同方法生成 7 个区域影像的可视化比较

Fig. 7　Visual comparison of 7 regional images generated using different methods

3.3　时空-文本联合交互引导可视化分析

为验证时空交互引导分支与文本交互引导分支的解耦性与协同性，本文开展了控制变量可视化试验，结果如图 8 所示。当保持文本提示近似而改变时空信息时（图 8(a)），本文方法能在维持地物语义内容一致的同时，准确生成符合相应时空背景约束的视觉纹理与整体风格，如新疆乡村变为湖南乡村、东北夏季农田变为冬季农田。相反，对比方法无法确保纹理和风格与相应的时空背景保持一致，并且在时空信息变化时无法保持语义内容一致性。当保持时空信息近似而改变文本提示时（图 8(b)），本文方法能够准确响应文本语义的变化，且地物的纹理与风格保持一致。而对比方法在文本提示改变时，难以维持稳定的地物纹理和风格。上述现象的原因在于，对比方法采用特征拼接或相加的方式进行多模态特征的交互引导，缺乏解耦的特征交互空间，导致稀疏的时空特征容易被稠密的文本语义特征所干扰甚至抑制。而本文方法通过新增独立并行的时空交互引导分支，为时空特征提供了独立的特征交互空间，从根本上避免了文本特征与时空特征之间的相互干扰，实现了二者的解耦与协同引导。



(a) 保持文本提示近似一致，使用不同的时空信息　　(b) 保持时空信息近似一致，使用不同的文本提示

图 8　时空信息和文本提示对生成影像的影响

Fig. 8　The effects of spatio-temporal information and text prompt in generating images

4　讨　论

4.1　独立投影模块对生成结果可控性的影响

仅通过异构频率编码难以实现对多源时空信息的语义解耦。为此，本文为每种时空信息分别设计了独立的轻量化投影模块。消融试验（表 2）表明，在去除独立投影模块后，生成影像的各项评估指标性能均出现不同程度的下降。具体而言，衡量时空分布一致性的 FID 指标上升至 4.860 4，衡量结构纹理一致性的 MS-SSIM 指标下降至 0.171 6，衡量语义内容一致性的 CLIP Score 指标亦有微弱下降。这些结果表明，只使用异构频率编码处理多

源时空信息，会导致特征空间中的语义混淆，使得模型难以准确解耦不同类型时空信息的独特语义。如图 9 所示，本文设计的独立投影模块，能够将不同类型的时空信息映射为解耦的特征表示。这为扩散模型提供了无歧义的引导信号，从而显著提升了生成结果的时空分布一致性和结构纹理一致性。

表 2 针对独立投影模块的消融试验定量结果

Tab. 2 Quantitative results of the ablation experiment for the independent projection module

模块	FID	MS-SSIM	CLIP Score
不使用独立投影模块	4.860 4	0.171 6	25.770 2
使用独立投影模块	3.457 2	0.225 8	25.770 5

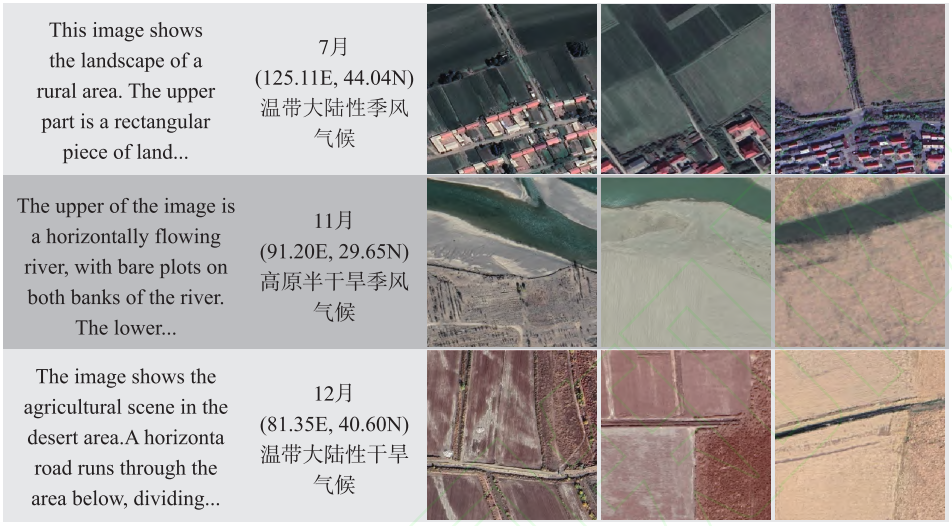


图 9 独立投影模块对生成结果影响的可视化比较

Fig. 9 Visual comparison of the impact of independent projection modules on the generated results

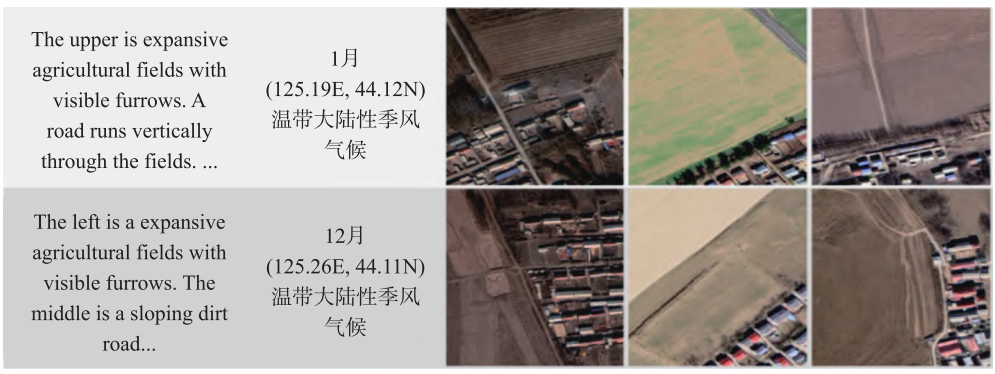
4.2 异构频率编码对生成结果可控性的影响

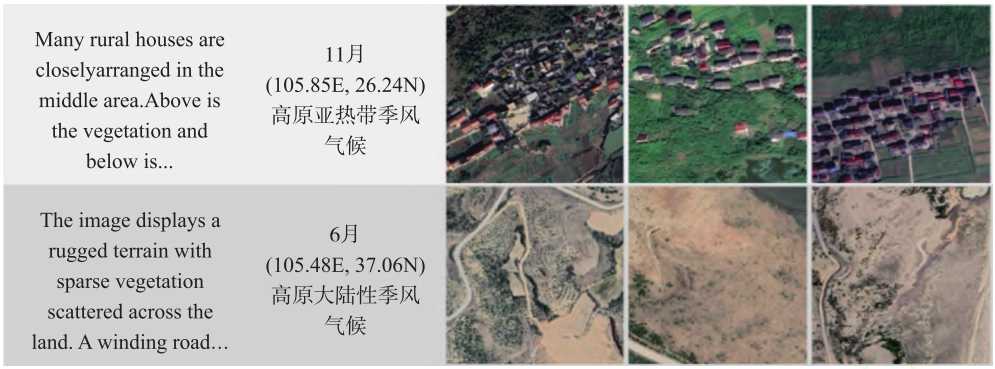
为实现对多源时空信息差异化属性的准确表征，本文提出了基于差异化维度的异构频率编码策略。消融试验（表 3）表明，将异构频率编码替换为标准的 MLP 投影后，生成影像的各项评估指标性能均出现不同程度的下降。具体而言，衡量时空分布一致性的 FID 指标性能下降了 31.05%，衡量结构纹理一致性 MS-SSIM 指标性能下降了 6.33%，衡量语义内容一致性的 CLIP Score 指标亦有微弱下降。这些结果表明了标准 MLP 投影处理多源时空信息的局限性。其统一的编码方式难以兼顾对多源时空信息差异化属性的准确表示，进而导致扩散模型无法准确建立时空特征与地物视觉特征之间的映射关系。相反，如图 10 所示，本文提出的异构频率编码策略能够有效兼顾多源时空信息的差异化属性。这为生成过程提供了准确的时空特征引导信号，从而显著提升了生成结果的时空分布一致性和结构纹理一致性。

表 3 针对异构频率编码策略的消融试验定量结果

Tab. 3 Quantitative results of the ablation experiment for heterogeneous frequency coding strategy

编码策略	FID	MS-SSIM	CLIP Score
标准 MLP 编码	4.530 6	0.211 5	25.736 5
异构频率编码	3.457 2	0.225 8	25.770 5





(a) 文本提示　(b) 时空信息　(c) 真实参考影像　(d) 标准MLP投影　(e) 异构频率编码

图 10　异构频率编码策略对生成结果影响的可视化比较

Fig. 10　Visualization comparison of the impact of heterogeneous frequency encoding strategy on the generated results

4.3　解耦注意力机制对生成结果可控性的影响

为避免时空特征与文本特征在交互过程中的相互干扰，本文提出了基于解耦注意力的时空-文本联合交互引导机制。为验证其有效性，本文开展消融试验，在保持扩散模型原有文本交互引导分支的基础上，采用标准交叉注意力建模时空特征与视觉特征的交互（图 11）。

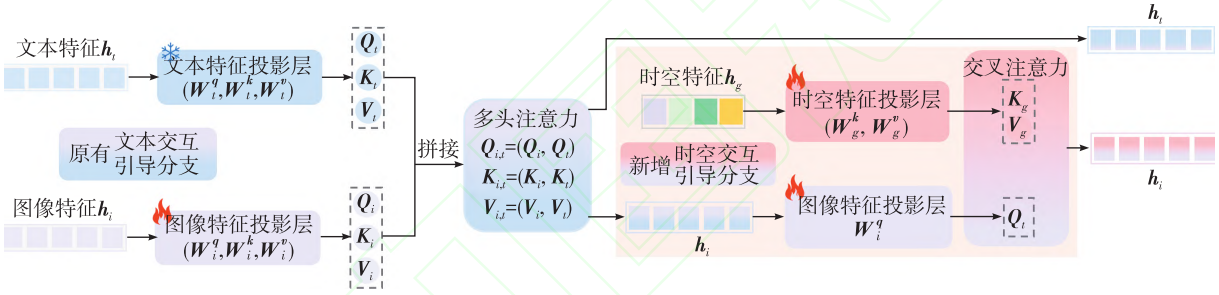


图 11　基于标准交叉注意力的时空特征与视觉特征交互引导

Fig. 11　Guided interaction by spatio-temporal features and visual features based on the standard cross-attention

试验结果见表 4，当采用标准交叉注意力时，生成结果的各项指标均出现性能下降。FID 指标从 3.457 2 弱化至 7.848 5，MS-SSIM 指标下降了 5.53%，CLIP Score 亦下降了 1.65%。直接利用标准交叉注意力建模时空特征与视觉特征的交互，会对视觉特征中已建立的文本语义进行强制重构。这迫使模型在适应时空特征对地物纹理和风格施加约束的同时，破坏了由文本交互引导分支确立的地物语义内容特征，最终导致生成影像在时空分布与语义内容一致性上的双重退化。相反，本文提出的基于解耦注意力的时空-文本联合交互引导机制通过构建独立并行的时空交互引导分支，使得文本特征与时空特征能够在互不干扰的前提下分别与视觉特征进行深度交互。如图 12 所示，标准交叉注意力生成的影像出现了严重的语义结构崩塌与纹理失真；而采用本文机制生成的影像不仅结构清晰、语义准确，更精准还原了特定时空背景下的视觉纹理。这充分证实了解耦交互机制在多模态特征协同引导上的显著优势。

表 4　针对解耦注意力机制的消融试验定量结果

Tab. 4　Quantitative results of the ablation experiment for the decoupled attention mechanism

注意力机制	FID	MS-SSIM	CLIP Score
标准交叉注意力	7.848 5	0.213 3	25.345 5
解耦注意力	3.457 2	0.225 8	25.770 5

4.4　微调适配秩的影响

微调适配秩是影响模型计算成本与数据分布拟合能力的关键参数。为探究不同的适配秩对生成结果的影响，同时兼顾生成质量与计算成本的平衡，本文在 128～1024 的范围内对适配秩进行了消融试验。如表 5 所示，随着适配秩从 128 增至 1024，生成影像质量呈持续优化趋势。FID 指标从 8.161 4 持续下降至 3.457 2；MS-SSIM 指标从 0.186 1 提升至 0.225 8；衡量语义内容一致性的 CLIP-Score 指标亦从 25.564 0 上升至 25.770 5。

在较低的适配秩设置下，模型的特征表达能力有限，不足以有效建模遥感影像中复杂的纹理细节与时空分布规律，呈现出明显的欠拟合现象。当适配秩增加至1024时，各项指标均达到最优且未出现过拟合带来的性能退化。

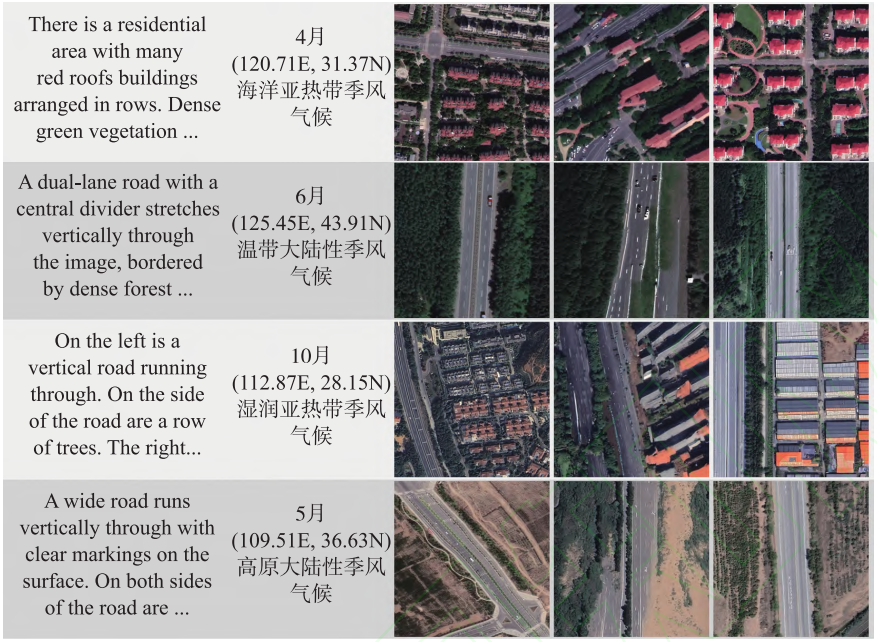


图12 基于解耦注意力的特征交互引导机制对生成结果影响的可视化比较

Fig. 12 Visualization comparison of the impact of the interaction guidance mechanism between features based on decoupled attention

表5 针对适配秩的消融试验定量结果

Tab. 5 Quantitative results of the ablation experiment for the adaptation rank

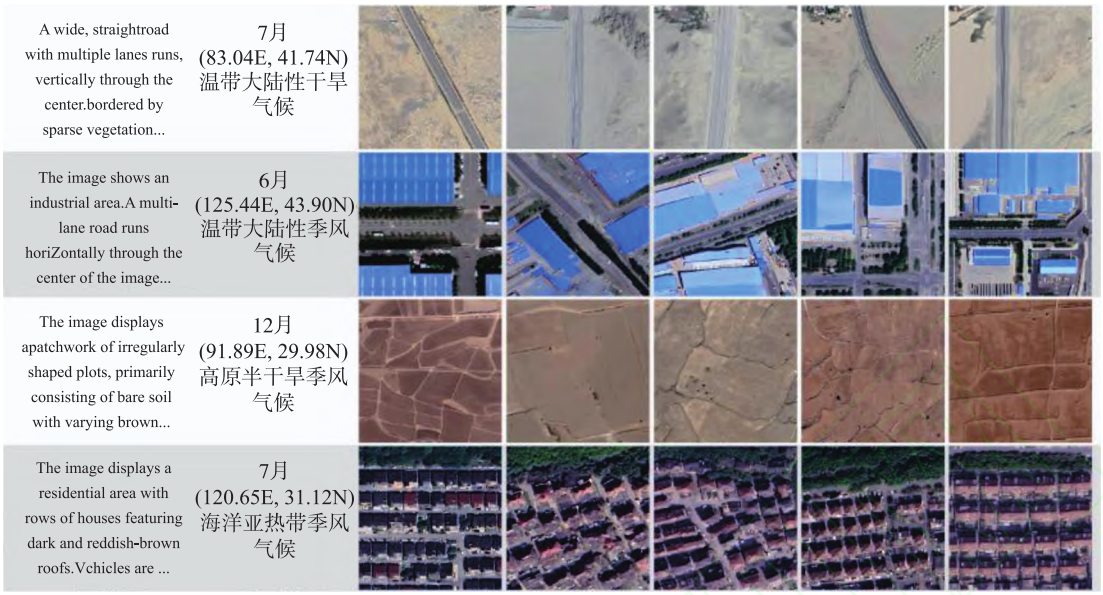
适配秩	FID	MS-SSIM	CLIP-Score
128	8.161 4	0.186 1	25.564 0
256	6.567 2	0.203 8	25.746 0
512	4.574 8	0.224 8	25.767 0
1024	3.457 2	0.225 8	25.770 5

图13展示了不同适配秩设置下的生成结果。在较低的适配秩设置下，生成影像表现出显著的结构模糊与纹理丢失，如相邻建筑的边界不清晰，这反映了模型对影像细节纹理的建模能力存在显著缺陷。相比之下，当适配秩提升至1024时，模型生成影像的几何结构清晰，并且纹理细节真实。综上所述，为捕捉遥感影像的高频纹理细节和复杂的时空分布规律，设置较大的适配秩是必要的。考虑到生成质量与计算效率的平衡，本文最终选取1024作为最优的适配秩设置。

4.5 分割掩码对文本描述构建的影响

为验证地物要素分割结果对视觉语言模型生成文本描述的影响，本文进行了定性对比。针对同一张影像，分别构建准确分割掩码（人工提示修正）与不准确掩码（全自动提示），随后将它们分别与原始影像和思维链提示共同输入视觉语言模型，以比较生成的文本描述在地物要素类别与布局上是否存在显著差异。

如图14所示，尽管输入的分割掩码存在差异，但视觉语言模型生成的文本描述在关键的地物类别和布局上保持一致。这主要得益于本文采取的多源信息互补与层次化思维链推理机制。首先，模型以真实的原始影像为视觉基准，掩码仅作为粗粒度布局参考。其次，思维链提示中明确包含“作为参考的分割掩码图可能是错误的，最终输出必须包含遥感影像中实际可见的内容”这一式约束，强制模型在掩码与影像冲突时优先信赖原始影像。最后，视觉语言模型自身具备强大的视觉理解能力，能够在思维链的提示下对影像进行层次化的理解与分析。因此，文本描述在语义层面的稳定性，确保了最终生成影像不受分割结果的影响。



(a) 文本提示　(b) 时空信息　(c) 真实参考影像　(d) 适配秩128　(e) 适配秩256　(f) 适配秩512　(g) 适配秩1024

图 13　适配秩对生成结果影响的可视化比较

Fig. 13　Visualization comparison of the impact of adaptation rank on the generated results

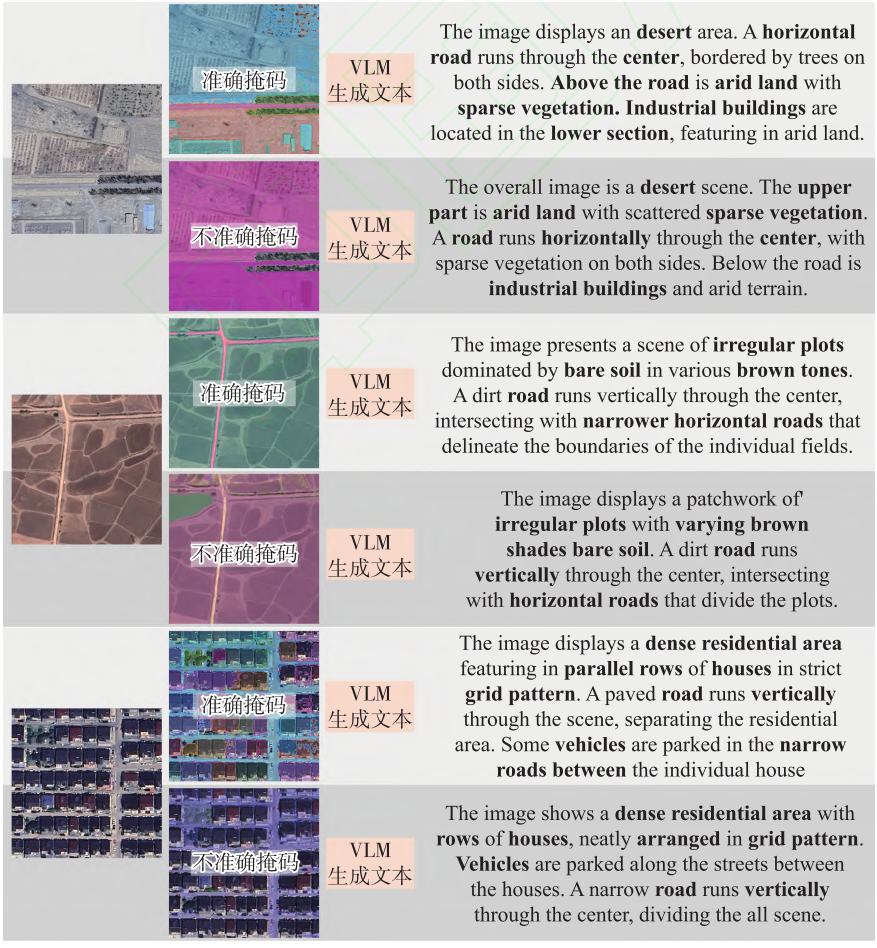


图 14　分割掩码对于视觉语言模型生成文本描述的影响

Fig. 14　The influence of segmentation masks on the generation of text descriptions by visual language model

5 结 论

针对当前高分光学遥感影像生成模型缺乏时空背景有效控制的问题，本文提出了一种面向时空可控的影像生成方法。通过设计异构频率编码策略与基于解耦注意力的时空-文本联合交互引导机制，有效解决了多源时空信息的语义解耦与多模态特征干扰问题。在全国7个典型区域的数据集上的试验表明，本文方法在时空分布一致性、结构纹理一致性和语义内容一致性3个维度上均显著优于现有主流方法，能够生成逼真且符合特定时空背景的高分光学遥感影像。未来研究将在现有基础上进一步拓展。一是探索更高分辨率的生成模型以实现小尺寸物体的生成。二是拓展数据集规模以实现全球尺度下的时空可控生成。此外，将探索多光谱与合成孔径雷达影像的生成，为遥感智能解译系统提供更全面的动态仿真环境。

参考文献

[1] HUANG Wei, CUI Zhimei, HUANG Zhidu, et al. Research on building extraction based on object-oriented CART classification algorithm and GF-2 Satellite images[J]. Journal of Geodesy and Geoinformation Science, 2024, 7(4): 5-18.

[2] ZHAO Bofei, SUI Haigang, ZHU Yihao, et al. Real-time rescue target detection based on UAV imagery for flood emergency response [J]. Journal of Geodesy and Geoinformation Science, 2024, 7(1): 74-89.

[3] HAN Zheng, LING Ziyan, DONG Li, et al. Heterogeneity effect of human disturbances on landscape patterns in the Yellow River Delta wetland, China[J]. Journal of Geodesy and Geoinformation Science, 2024, 7(4): 75-93.

[4] 杨元喜. 地理空间数字孪生与时空智能[J]. 测绘学报, 2025, 54(2): 213-220. DOI: 10.11947/j.AGCS.2025.20240515.
YANG Yuanxi. Digital twin and spatio-temporal intelligence of geospatial information system[J]. Acta Geodaetica et Cartographica Sinica, 2025, 54(2): 213-220. DOI: 10.11947/j.AGCS.2025.20240515.

[5] 李德仁, 王密, 肖晶, 等. 论无所不在的时空智能[J]. 遥感学报, 2025, 29(6): 1388-1398.
LI Deren, WANG Mi, XIAO Jing, et al. On ubiquitous spatio-temporal intelligence[J]. National Remote Sensing Bulletin, 2025, 29(6): 1388-1398.

[6] 陈军, 艾廷华, 闫利, 等. 智能化测绘的混合计算范式与方法研究[J]. 测绘学报, 2024, 53(6): 985-998. DOI: 10.11947/j.AGCS.2024.20240131.
CHEN Jun, AI Tinghua, YAN Li, et al. Hybrid computational paradigm and methods for intelligentized surveying and mapping[J]. Acta Geodaetica et Cartographica Sinica, 2024, 53(6): 985-998. DOI: 10.11947/j.AGCS.2024.20240131.

[7] TAO Chao, QI Ji, ZHANG Guo, et al. TOV: the original vision model for optical remote sensing image understanding via self-supervised learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 4916-4930.

[8] 龚健雅, 许越, 胡翔云, 等. 遥感影像智能解译样本库现状与研究[J]. 测绘学报, 2021, 50(8): 1013-1022. DOI: 10.11947/j.AGCS.2021.20210085.
GONG Jianya, XU Yue, HU Xiangyun, et al. Status analysis and research of sample database for intelligent interpretation of remote sensing image[J]. Acta Geodaetica et Cartographica Sinica, 2021, 50(8): 1013-1022. DOI: 10.11947/j.AGCS.2021.20210085.

[9] ZHENG Z, ERMON S, KIM D, et al. Changen2: multi-temporal remote sensing generative change foundation model[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(2): 725-741.

[10] ZAN Yujie, JI Shunping, CHAO Songtao, et al. Open-vocabulary generative vision-language models for creating a large-scale remote sensing change detection dataset[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2025, 225: 275-290.

[11] ZHENG Zhuo, TIAN Shiqi, MA Ailong, et al. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2024: 21761-21770.

[12] XU Yonghao, YU Weikang, GHAMISI P, et al. Txt2Img-MHN: remote sensing image generation from text using modern Hopfield networks[J]. IEEE Transactions on Image Processing, 2023, 32: 5737-5750.

[13] ZHAO Rui, SHI Zhenwei. Text-to-remote-sensing-image generation with structured generative adversarial networks[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 8010005.

[14] LIU Yidan, YUE Jun, XIA Shaobo, et al. Diffusion models meet remote sensing: principles, methods, and perspectives[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 4708322.

[15] 张新长, 赵元, 齐霁, 等. 基于AI大模型的文生图技术方法研究及应用[J]. 地球信息科学学报, 2025, 27(1): 10-26.
ZHANG Xinchang, ZHAO Yuan, QI Ji, et al. Research and application of text-to-image technology based on AI foundation models[J]. Journal of Geo-information Science, 2025, 27(1): 10-26.

[16] YUAN Zhiqiang, HAO Chongyang, ZHOU Ruixue, et al. Efficient and controllable remote sensing fake sample generation based on diffusion model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5615012.

[17] XU Yue, LIU Honghao, YANG Ruixia, et al. Remote sensing image semantic segmentation sample generation using a decoupled latent

diffusion framework[J]. Remote Sensing，2025，17(13)：2143.

[18] BAGHIRLI O，ASKAROV H，IBRAHIMLI I，et al. SatDM：synthesizing realistic satellite image with semantic layout conditioning using diffusion models[EB/OL]. [2025-11-03]. http://arxiv.org/abs/2309.16812.

[19] DONG Runmin，YUAN Shuai，FENG Litong，et al. Transferable image synthesis for remote sensing semantic segmentation via joint reference-semantic fusion[J]. Information Fusion，2026，127：103839.

[20] TANG Datao，CAO Xiangyong，WU Xuan，et al. AeroGen：enhancing remote sensing object detection with diffusion-driven data generation[C]//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville：IEEE，2025：3614-3624.

[21] ZHANG Mu，LIU Yunfan，LIU Yue，et al. CC-diff：spatially controllable text-to-image synthesis for remote sensing with enhanced contextual coherence[J]. IEEE Transactions on Geoscience and Remote Sensing，2025，63：5645316.

[22] OU Ruizhe，YAN Haotian，WU Ming，et al. A method of efficient synthesizing post-disaster remote sensing image with diffusion model and LLM[C]//Proceedings of 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. [S. l.]：IEEE，2023：1549-1555.

[23] SEBAQ A，ELHELW M. RSDiff：remote sensing image generation from text using diffusion model[J]. Neural Computing and Applications，2024，36(36)：23103-23111.

[24] LIU Chenyang，CHEN Keyan，ZHAO Rui，et al. Text2Earth：unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model[J]. IEEE Geoscience and Remote Sensing Magazine，2025，13(3)：238-259.

[25] YU Zhiping，LIU Chenyang，LIU Liqin，et al. MetaEarth：a generative foundation model for global-scale remote sensing image generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2025，47(3)：1764-1781.

[26] KHANNA S，LIU P，ZHOU Linqi，et al. DiffusionSat：a generative foundation model for satellite imagery[EB/OL]. [2025-11-03]. https://arxiv.org/abs/2312.03606.

[27] BAI Jinze，BAI Shuai，YANG Shusheng，et al. Qwen-VL：a versatile vision-language model for understanding，localization，text reading，and beyond[EB/OL]. [2025-11-03]. http://arxiv.org/abs/2308.12966.

[28] LIU Shilong，ZENG Zhaoyang，REN Tianhe，et al. Grounding DINO：marrying DINO with grounded pre-training for open-set object detection[C]//Proceedings of Computer Vision-ECCV 2024. Cham：Springer，2025：38-55.

[29] REN Tianhe，LIU Shilong，ZENG Ailing，et al. Grounded SAM：assembling open-world models for diverse visual tasks[EB/OL]. [2025-11-03]. http://arxiv.org/abs/2401.14159.

[30] PEEBLES William，XIE Saining. Scalable diffusion models with transformers[C]//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris：IEEE，2023：4172-4182.

[31] HU E，SHEN Y，WALLIS P，et al. LoRA：low-rank adaptation of large language models[C]//Proceedings of 2022 International Conference on Learning Representations. San Diego：OpenReview.net，2022.

[32] TANG Datao，CAO Xiangyong，HOU Xingsong，et al. CRS-diff：controllable remote sensing image generation with diffusion model[J]. IEEE Transactions on Geoscience and Remote Sensing，2024，62：5638714.

[33] HEUSEL M，RAMSAUER H，UNTERTHINER T，et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[C]//Proceedings of 2017 Neural Information Processing Systems. Long Beach：Curran Associates，Inc.，2017：6629-6640.

[34] WANG Z，SIMONCELLI E P，BOVIK A C. Multiscale structural similarity for image quality assessment[C]//Proceedings of 2003 Asilomar Conference on Signals，Systems & Computers. Pacific Grove：IEEE，2003：1398-1402.

[35] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]//Proceedings of 2021 International Conference on Machine Learning. San Diego：PMLR，2021：8748-8763.

（责任编辑：张　琳）

Controllable generation of high-resolution optical remote sensing image explicitly guided by spatio-temporal information

SHI Tiandong[1], ZHAO Ling[1], ZHAO Wenhao[2], QI Ji[3], CUI Hao[4], PENG Chengli[1], ZHANG Xinchang[3]

1. *School of Geosciences and Info-Physics, Central South University, Changsha* 410083, *China*;
2. *National Geomatics Center of China, Beijing* 100830, *China*;
3. *School of Geography and Remote Sensing, Guangzhou University, Guangzhou* 510006, *China*;
4. *State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan* 430079, *China*

Abstract: The visual appearance of land cover objects in high-resolution optical remote sensing images is significantly influenced by seasonal evolution and regional differences. Enhancing the spatio-temporal controllability of generation models to reproduce object features under specific spatio-temporal contexts accurately remains a critical challenge. Existing research has limitations in the encoding method of multi-source spatio-temporal information, as well as the interaction guidance method between encoded spatio-temporal features and visual features, making it difficult to accurately model the precise mapping between spatio-temporal conditions and the visual appearance of land cover objects. To address this problem, this paper proposes a framework for spatio-temporal controllable high-resolution optical remote sensing image generation. First, a multi-source spatio-temporal information encoding strategy considering attribute differences is designed, which utilizes heterogeneous frequency encoding and independent projections to transform diverse spatio-temporal information into accurate and decoupled feature representations, thereby modeling the unique properties of diverse spatio-temporal information. Second, an interaction guidance mechanism between spatio-temporal features and visual features based on decoupled attention is designed. This mechanism employs an independent parallel attention branch to facilitate deep interaction between spatio-temporal features and visual features, effectively leveraging the constraining role of spatio-temporal information without interfering with text-guided generation. We adopt low-rank adaptation to efficiently transfer domain knowledge by optimizing only low-rank decomposition matrices, thereby preserving the pre-trained generative priors of the base model. Experiments on a large-scale dataset covering seven typical regions in China demonstrate that the proposed method outperforms state-of-the-art methods by 46.69% and 14.67% in terms of spatio-temporal distribution consistency and structural-textural consistency, respectively. These results confirm the controllability and generalization potential of the proposed framework across diverse spatio-temporal scenarios.

Key words: spatio-temporal intelligence; remote sensing image; generation model; diffusion model; deep learning; spatio-temporal information

Foundation support: The National Natural Science Foundation of China (Nos. 42571533; 42371406); Funded by State Key Laboratory of Spatial Datum(No. SKLSD2026-KF-27)

First author: SHI Tiandong (1996—), male, PhD candidate, majors in remote sensing intelligent interpretation and controllable generation.
E-mail: csushitd@csu.edu.cn
Corresponding author: QI Ji
E-mail: jameschi95@foxmail.com