Bridging Data Gaps in Open-Pit Mining: A Factor-Isolated Framework for Long-Tail Scenario Generation

Ruixi Qiao, Chenglong Ye, Jie Cheng, Peizhe Gong, Xingyuan Dai, Xiaoyan Gong, and Yisheng Lv

***Abstract*—The development of reliable autonomous driving systems for mining is critically limited by the scarcity of long-tail scenario data. While generative models provide a promising direction, they often lack the fine-grained control and domain-specific realism required in safety-critical applications. To address this challenge, we propose a decoupled training pipeline that models a scene's fundamental components, namely style, layout, and motion, as architecturally separated and controllable factors. Firstly, style and layout are jointly fine-tuned to capture the unique textures and spatial structures of mining environments, ensuring high-fidelity visuals and precise spatial control. Subsequently, we selectively fine-tune the temporal layers of the video generation model. This captures empirically plausible visual motion patterns of heavy machinery, creating short, high-fidelity videos. Experiments show that our pipeline not only outperforms general video baselines but also improves downstream task performance. By bridging the gap between static and dynamic data, our framework supports robust perception and supplies synthetic dynamic scenarios for vision-language-model (VLM)-based training and evaluation in autonomous mining.**

***Index Terms*—Autonomous driving, image generation, video generation, diffusion model**

I. Introduction

While autonomous driving technology holds the promise of revolutionizing the mining industry, its success depends on developing perception and decision-making models of high robustness. Unlike structured urban settings, open-pit mines are complex environments demanding access to vast and diverse training data that capture the full spectrum of operational scenarios.

A primary obstacle is the long-tail problem [1, 2]. Rare but critical events, such as equipment collisions, sudden rockfalls, or extreme weather, are sparsely represented in collected datasets, yet these are precisely the scenarios that can lead to catastrophic failures. Generative artificial intelligence (AI) has emerged as a powerful tool to synthesize this missing data, but existing methods often fail to generate high-quality mining scenarios. They fall short in three critical areas: (1) high-fidelity style: failing to capture the unique visual textures and atmospherics of a mine; (2) precise layout control: lacking the ability to place specific hazards, like a rockfall, in precise locations to create meaningful test cases; and (3) plausible motion dynamics: static generators miss critical motion signatures, while general video models often generate unrealistic movements.

A central challenge in adapting generative foundation models to mining scenarios is that the relevant factors of a dynamic scene are heterogeneous. Visual appearance is dominated by dust, illumination, road texture, and machinery surface details; spatial layout depends on object classes, scales, and relative positions on unstructured haul roads; and temporal dynamics are governed by the slow, high-inertia motion of heavy machinery. Monolithic fine-tuning entangles these factors in a single parameter update, making it difficult to control them independently and to diagnose which factor contributes to downstream improvements.

We therefore formulate mining scenario generation as a factor-isolated adaptation problem. Instead of treating fine-tuning merely as an engineering tool, we use it to impose a structured adaptation protocol: mining-specific appearance, layout grounding, and motion dynamics are assigned to architectural components that are most relevant to each factor. Specifically, the style factor is adapted through the cross-attention layers of an image generator [3–5]; the layout factor is adapted through the grounding module that maps object boxes into spatial conditions; and the motion factor is adapted through the temporal convolution and attention layers [6]. This design preserves the generative prior of the foundation models while enabling controllable specialization to unstructured mining environments.

Unlike urban driving generation methods that often rely on high-definition (HD) maps, lane topology, and traffic-light priors, our framework learns controllability from image-level attributes, object-level bounding boxes, and real mining video clips. This makes the method more suitable for open-pit mines, where roads are unstructured, scene geometry changes frequently, and rare safety-critical events are difficult to collect at scale.

The main contributions of our work are as follows:

Manuscript received: 15 May 2026; revised: 28 May 2026; accepted: 1 June 2026. *(Corresponding author: Ruixi Qiao.)*

Citation: R. Qiao, C. Ye, J. Cheng, P. Gong, X. Dai, X. Gong, and Y. Lv, Bridging data gaps in open-pit mining: A factor-isolated framework for long-tail scenario generation, *Int. J. Intell. Control Syst.*, 2026, 31(2), 205–213.

Ruixi Qiao, Chenglong Ye, Jie Cheng, Peizhe Gong, Xingyuan Dai, Xiaoyan Gong, and Yisheng Lv are with State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: qiaoruixi2024@ia.ac.cn; yechenglong2024@ia.ac.cn; chengjie2022@ia.ac.cn; gongpeizhe21@mails.ucas.ac.cn; xingyuan.dai@ia.ac.cn; xiaoyan.gong@ia.ac.cn; yisheng.lv@ia.ac.cn).

Digital Object Identifier 10.62678/IJICS202606.10375

(1) We propose a factor-isolated adaptation framework that separates mining-specific style, layout, and motion into dedicated low-rank adaptation (LoRA) subspaces.

(2) We build a controllable long-tail mining generation pipeline using taxonomy-guided prompts, image-level attributes, object-level grounding, and real mining video clips.

(3) We validate the design with monolithic and partial-decoupling baselines, layout controllability, video quality metrics, object detection, and vision-language-model (VLM) decision-making.

II. Related Work

In this section, we review generative models for autonomous driving and video generation methods related to our work. We focus on the gap between existing urban driving generation methods and the requirements of unstructured open-pit mining scenarios, as well as the limitations of general video generation and adaptation methods for safety-critical motion synthesis.

A. Generative Models for Autonomous Driving

Data-driven scene synthesis has become an important paradigm for addressing data scarcity in autonomous driving. Recent works such as MagicDrive [7] and DriveDreamer-2 [8] demonstrate that diffusion models can generate diverse driving scenes with explicit geometric or semantic control. Panacea [9] further explores panoramic and controllable video generation, while Vista [10] studies high-fidelity and generalizable driving world models. These methods show that generative models can provide scalable data for perception and planning research.

However, most existing driving generation methods are designed around urban traffic assumptions. Their controllability often depends on structured priors such as HD maps, lane topology, traffic lights, road boundaries, and well-defined agent categories. These assumptions do not hold in open-pit mines, where haul roads are unstructured, lane markings are absent, road geometry changes over time, and scene appearance is strongly affected by dust, snow, mud, fog, illumination, and sensor contamination. Moreover, many mining hazards, such as rockfalls, large puddles, stopped haul trucks, and auxiliary equipment intrusions, are rare and difficult to collect under controlled safety conditions.

Our work differs from urban driving generation in both task formulation and control signals. Instead of relying on map-based control, we use image-level mining attributes and object-level bounding boxes to specify long-tail mining scenarios. This enables controllable generation in environments where HD maps or lane-level topology are unavailable. In addition, our evaluation emphasizes downstream utility for autonomous mining, including object detection and VLM decision-making.

B. Video Generation and World Models

General text-to-video models have advanced rapidly with architectures such as stable video diffusion [11], Sora [12], and CogVideoX [13]. These models benefit from large-scale pre-training and can generate visually impressive videos from text prompts. However, their general-domain priors are not sufficient for open-pit mining scenarios. When directly prompted with mining scenes, they may hallucinate unrealistic road structures, produce inconsistent machinery geometry, or generate vehicle motion that appears to slide across frames rather than maintaining visually consistent motion.

Parameter-efficient adaptation methods provide a practical way to specialize large generative models. LoRA [4] reduces the number of trainable parameters by learning low-rank updates while keeping the original model weights frozen. AnimateDiff [14] demonstrates that temporal modules can extend personalized text-to-image models to video generation. Recent work also suggests that naively applying LoRA to video tasks may be insufficient for maintaining temporal continuity, motivating more structured temporal conditioning or adaptation strategies [15]. These observations indicate that adaptation for video generation should consider where and how temporal information is injected.

Our framework follows this direction but targets a different problem: controllable long-tail scenario generation for unstructured mining environments. Rather than applying a single adapter to the whole generation pipeline, we isolate appearance, layout, and motion into different adaptation subspaces. The style adapter specializes the image generator in mining textures and atmospheric effects, the layout adapter improves grounding-based object placement, and the motion adapter modifies only temporal layers to capture empirical motion patterns from real mining videos. This factor-isolated design reduces interference between static appearance and dynamic motion and provides a clearer basis for controlled ablation against monolithic fine-tuning.

III. Preliminary

This section outlines the foundational technologies leveraged in our pipeline: the Stable Diffusion XL (SDXL) text-to-image model, the LoRA fine-tuning technique, and the Hotshot-XL text-to-video framework.

A. Stable Diffusion

Stable Diffusion is a latent diffusion model that generates images by reversing a diffusion process in a computationally efficient, low-dimensional latent space [3]. The core of this process is a UNet model trained to predict and remove noise from a latent representation. This denoising is guided by text prompts, which are encoded and injected into the UNet via cross-attention mechanisms.

We utilize SDXL, a superior version featuring a larger UNet backbone and a dual text-encoder system. This enhanced architecture provides a more nuanced understanding of prompts, resulting in significantly higher-fidelity and more detailed images.

B. Low-Rank Adaptation

To adapt large models like SDXL efficiently, we employ LoRA, a parameter-efficient fine-tuning technique [4]. LoRA operates on the hypothesis that the change in weights during adaptation has a low intrinsic rank. Instead of updating the original pre-trained weights W_0, LoRA freezes them and

injects a pair of small, trainable low-rank matrices (B and A). The model's forward pass is modified as

$$h = W_0 x + \frac{\alpha}{r} BAx \tag{1}$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d' \times r}$ are trainable low-rank matrices, r is the LoRA rank, and α is a scaling coefficient. By training only A and B, LoRA drastically reduces the number of trainable parameters, enabling cost-effective domain specialization while retaining the base model's learned priors.

C. Hotshot-XL

Hotshot-XL is an open-source framework that extends SDXL's capabilities from image to video synthesis [6]. It achieves this by integrating specialized temporal layers directly into the SDXL UNet architecture. While the original layers process spatial information within a frame, these new temporal layers create connections across frames. This mechanism is crucial for maintaining temporal consistency and generating coherent motion.

IV. Methodology

Our proposed data generation pipeline is built on a factor-isolated adaptation principle. Rather than updating a generative foundation model with a single monolithic adapter, we explicitly decompose a mining scene into three controllable factors: visual style, spatial layout, and temporal motion. Each factor is assigned to a dedicated adaptation subspace that is aligned with its semantic role in the generation process. Fig. 1 illustrates the overall architecture, which consists of three components: long-tail scenario definition and data preparation, static image adaptation for style and layout control, and video adaptation for mining-specific motion dynamics.

A. Problem Formulation

Let $\mathcal{D}_{\text{img}} = \{(x_i, c_i, b_i)\}_{i=1}^{N}$ denote the image-level training set, where x_i is a mining image, c_i is the corresponding textual or attribute-level description, and $b_i = \{(o_{ij}, r_{ij})\}_{j=1}^{M_i}$ represents object-level grounding annotations. Here, o_{ij} denotes the object category and r_{ij} denotes its bounding box. Let $\mathcal{D}_{\text{vid}} = \{(v_k, c_k)\}_{k=1}^{K}$ denote the video-level training set, where $v_k = \{x_k^1, x_k^2, \ldots, x_k^T\}$ is a temporally contiguous mining clip with T frames.

Our goal is to learn a controllable conditional generator that synthesizes either static mining images or short video sequences from textual descriptions and, when available, spatial grounding constraints. We denote the adapted conditional distribution as

$$y \sim p_{\Theta,\Phi}(y \mid c, b) \tag{2}$$

where y denotes a generated image or video clip, c denotes the textual or attribute-level condition, b denotes the object-level grounding box, Θ denotes the frozen foundation-model parameter, and Φ denotes the trainable adapter parameter.

We conceptually decompose a dynamic mining scene into three factors: visual style s, spatial layout l, and temporal motion m. These factors are not directly supervised as independent labels. Instead, we implement this decomposition architecturally by assigning each factor to a dedicated set of LoRA parameters

$$\Phi = \{\phi_s, \phi_l, \phi_m\} \tag{3}$$

where ϕ_s is the style adapter, ϕ_l is the layout adapter, and ϕ_m is the motion adapter. The pre-trained foundation parameters remain frozen during adaptation, and only the corresponding low-rank updates are optimized.

This formulation is designed to avoid the entanglement caused by monolithic fine-tuning. The style adapter learns mining-specific textures, lighting, and atmospheric effects; the layout adapter learns how mining objects and hazards should follow user-specified bounding boxes; and the motion adapter learns temporal changes from real mining clips without overwriting the learned static appearance and layout priors.

B. Long-Tail Scenario Definition and Data Preparation

A precise definition of long-tail scenarios is fundamental to our work. We collaborated with domain experts to categorize these events, as summarized in Table 1.

The training data consist of real AutoMine [16] samples and a small synthetic seed set. From AutoMine, we curated 14,250 real static frames for mining appearance and layout learning.

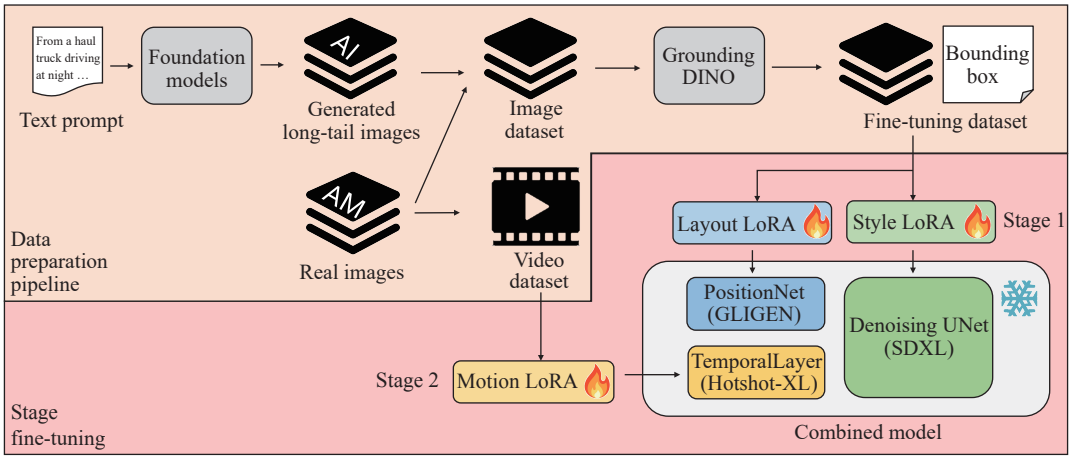


Fig. 1 Overview of the proposed factor-isolated generation pipeline. Real AutoMine data and synthetic seed images are used for static style/layout adaptation, followed by temporal adaptation with a motion LoRA for mining video generation.

Table 1 Taxonomy and occurrence statistics of long-tail scenarios in autonomous mining. The real-data proportions are computed from the curated AutoMine samples used in this study, while the generated proportions are computed from the 750 generated frames. Since a frame may contain multiple abnormal attributes, the generated proportions do not necessarily sum to 100%.

Category	Included scenario types	Occurrence in real training data	Coverage in generated training data
Abnormal weather	Foggy, blizzard, dust storms, heavy rain, and low-visibility weather conditions	176/14,250 samples, 1.2%	381/750 frames, 50.8%
Abnormal lighting	Night, strong backlight, low illumination, and camera obstruction caused by mud or dust	675/14,250 samples, 4.7%	267/750 frames, 35.6%
Abnormal events	Collision risk, abnormal intrusion by personnel, rockfall, slope collapse, and obstacles	81/14,250 samples, 0.6%	184/750 frames, 24.5%
Abnormal surfaces	Large puddles, debris or rocks on the road, icy patches, rutting, and wet or muddy road surfaces	967/14,250 samples, 6.8%	349/750 frames, 46.5%
Abnormal objects	Stopped vehicles, slow-moving haul trucks, auxiliary equipment, warning signs, and unusual objects	163/14,250 samples, 1.1%	134/750 frames, 17.9%

In addition, we generated 750 synthetic seed images using large language and vision-language models [17–19] to increase long-tail coverage, accounting for 5.0% of the final 15,000-image training set. This small ratio preserves the real AutoMine distribution while exposing the model to rare visual attributes.

Subsequently, we employed a multi-layered, semi-automated annotation strategy for the entire image dataset. For high-level scene description, the Qwen2.5-VL [17] model was used to tag each image with descriptive keywords, with a particular focus on long-tail attributes such as abnormal weather, lighting, road-surface hazards, stopped vehicles, and abnormal intrusions. To provide the object-level grounding required for fine-tuning the GLIGEN [5] component, we additionally employed Grounding DINO [20] to generate bounding box annotations for key objects, including mining trucks, excavators, people, warning signs, and road-surface hazards.

To ensure annotation quality, all automatically generated image-level labels and object-level boxes were manually reviewed. Samples with severe blur, ambiguous visibility, inconsistent hazard labels, or unreliable object localization were removed or corrected. The training and evaluation splits were separated at the clip or scenario level to avoid overlap between training prompts, real training clips, generated sequences, and held-out evaluation samples.

C. Stage 1: Static Adaptation for Style and Layout Control

The first stage aims to adapt a layout-aware text-to-image diffusion model to open-pit mining scenes. We use GLIGEN-SDXL as the static generation backbone because it supports text-conditioned synthesis with object-level grounding constraints. The pre-trained model parameters are frozen, and two LoRA modules are inserted into different architectural components.

The style LoRA ϕ_s is inserted into the cross-attention layers of the UNet. Its role is to adapt the visual distribution of the generator to mining-specific appearance, including dust, illumination changes, and unstructured road textures. The layout LoRA ϕ_l is inserted into the grounding module, including the PositionNet component that maps bounding boxes and object categories into spatial conditioning embeddings. Its role is to improve controllability over mining objects and long-tail hazards with unusual positions, scales, and configurations.

Given an image x_i, the encoder maps it into a latent representation z_i. Following the standard diffusion training objective, Gaussian noise ϵ is added at timestep t to obtain $z_{i,t}$. The static adaptation objective is

$$\min_{\phi_s,\phi_l} E_{(x_i,c_i,b_i)\sim\mathcal{D}_{\text{img}},t,\epsilon}\left[\left\|\epsilon-\epsilon_{\theta_0,\phi_s,\phi_l}(z_{i,t},t,c_i,b_i)\right\|_2^2\right] \tag{4}$$

where θ_0 denotes the frozen GLIGEN-SDXL backbone, and $\epsilon_{\theta_0,\phi_s,\phi_l}$ denotes the denoising network augmented with the style and layout adapters. During optimization, only ϕ_s and ϕ_l are updated.

This objective jointly learns domain appearance and spatial grounding while preserving their architectural separation. Compared with monolithic fine-tuning, this design allows the model to improve mining-style fidelity without degrading layout controllability, and to improve bounding-box compliance without forcing the entire image generator to relearn the visual distribution from limited mining data.

D. Stage 2: Motion Adaptation for Mining Video Generation

The second stage extends the adapted static generator to video generation. A visually realistic image is not sufficient for downstream VLM decision-making because safety-critical mining scenarios often depend on temporal cues, such as an approaching haul truck, a stopped vehicle on a slope, or a sudden intrusion into the driving corridor. However, directly fine-tuning all spatial and temporal layers of a video generator can overwrite the mining-specific appearance and layout priors learned in stage 1.

To avoid this problem, we first merge the trained style and layout LoRAs into the static backbone, obtaining a mining-specialized spatial generator denoted by θ_{SL}. We then construct a video UNet by initializing the spatial layers from θ_{SL} and importing the temporal convolution and attention layers from Hotshot-XL. During video adaptation, all spatial layers are frozen. A motion LoRA ϕ_m is inserted only into the temporal layers.

For a video clip $v_k = \{x_k^1, x_k^2, \ldots, x_k^T\}$, the frames are encoded into a latent sequence $Z_k = \{z_k^1, z_k^2, \ldots, z_k^T\}$. After adding Gaussian noise at timestep t, the video denoising objective is

$$\min_{\phi_m} E_{(v_k,c_k)\sim\mathcal{D}_{\text{vid}},t,\epsilon}\left[\left\|\epsilon-\epsilon_{\theta_{\text{SL}},\tau_0,\phi_m}(Z_{k,t},t,c_k)\right\|_2^2\right] \tag{5}$$

where τ_0 denotes the frozen pre-trained temporal layer from Hotshot-XL, and ϕ_m denotes the trainable temporal LoRA parameter. Only ϕ_m is optimized in this stage.

This stage isolates temporal learning from static appearance

learning. The model therefore learns mining-specific motion patterns from real video clips, such as slow acceleration, large-radius turning, and gradual relative scale changes of heavy machinery, while preserving the visual and spatial controllability acquired in stage 1.

V. Experiments

A. Implementation Details

Training: Style and layout LoRAs were trained on GLIGEN-SDXL with frozen backbone weights using 16 bit floating-point precision, 8 bit AdamW, learning rate 1×10^{-4}, batch size 1 per GPU, 4 gradient accumulation steps, and LoRA rank 32. For motion adaptation, we combined the adapted UNet with Hotshot-XL temporal layers and trained temporal LoRAs on 500 real mining video clips using 2 GPUs, effective batch size 8, learning rate 1×10^{-4}, and rank 16. Generated videos have a resolution of 512×288 and 8 frames.

All monolithic and partially decoupled variants in our experiments use the same training sets, prompts, sample counts, and generation protocol. The monolithic LoRA baseline jointly adapts all selected components without separating style, layout, and motion. Partial variants remove one factor-specific adapter while keeping the remaining protocol unchanged.

B. Results

We conducted a comprehensive evaluation of our pipeline through quantitative comparisons, qualitative assessments, and downstream task validation.

a. Quantitative Analysis

Our quantitative evaluation consists of two parts: an ablation study of the proposed factor-isolated adaptation strategy, and a comparison against representative general-purpose video generation baselines.

First, to validate the effectiveness of factor-isolated adaptation, we compare our full model with the original foundation model, incremental adapter variants, partially decoupled variants, and a monolithic LoRA baseline. We evaluate image quality using NIQE [21] and BRISQUE [22], and video quality using FVD [23]. All variants are evaluated using the same long-tail prompt set and generation protocol. This comparison is designed to determine whether the observed improvement comes from the proposed decoupling strategy rather than merely from additional training or increased model adaptation capacity.

To retain a compact evaluation of layout controllability without introducing a separate table, we report box accuracy at intersection over union (IoU) threshold 0.5 (BoxAcc@0.5) together with image and video quality metrics. For each generated image, the conditioning box is treated as the target box. We use Grounding DINO [20] as a fixed open-vocabulary evaluator to detect the corresponding object category in the generated image. For all methods, the same detector, prompts, confidence threshold, and non-maximum suppression settings are used. If multiple detections match the target class, the detected box with the highest IoU to the conditioning box is selected. BoxAcc@0.5 is defined as

$$\text{BoxAcc@0.5} = \frac{1}{N_b} \sum_{i=1}^{N_b} I\left[\text{IoU}(r_i, \hat{r}_i) > 0.5\right] \tag{6}$$

where N_b is the total number of evaluated grounding boxes, r_i is the conditioning box, and $\hat{r}_i$ is the detected box.

Table 2 reports the contribution of each factor-specific adapter and the cross-factor interference analysis. The first three rows show the incremental effect of static adaptation. Style adaptation improves NIQE from 7.68 to 6.85 and BRISQUE from 24.79 to 19.52, indicating improved mining-domain visual fidelity. Adding layout adaptation further improves NIQE to 6.42, BRISQUE to 16.81, and BoxAcc@0.5 from 41.6% to 58.1%, confirming better spatial grounding.

The comparison between "+ style and layout LoRA" and "ours: style + layout + motion" directly tests whether training the motion LoRA changes static image quality or layout controllability. The results show that adding the motion LoRA improves FVD from 480.5 to 413.8, while NIQE improves from 6.42 to 6.09, BRISQUE improves from 16.81 to 14.4, and BoxAcc@0.5 improves from 58.1% to 62.8%. This indicates that temporal adaptation improves video quality without degrading the static appearance and layout priors learned in stage 1. In contrast, the monolithic LoRA baseline jointly adapts the selected components without separating style, layout, and motion. Although it improves FVD to 446.4, its BoxAcc@0.5 drops to 49.7%, suggesting stronger interference with layout controllability.

The partial variants further show that the three adapters are

Table 2 Ablation and controlled comparison of factor-isolated adaptation. Style, layout, and motion denote adapters for cross-attention, grounding, and temporal layers, respectively. Bold marks the best metric values. "↑" and "↓" indicate that higher and lower values are better, respectively.

Model configuration	Adaptation design				Image/layout quality			Video quality
	Style	Layout	Motion	Isolation	NIQE ↓	BRISQUE ↓	BoxAcc@0.5 (%) ↑	FVD ↓
Baseline (SDXL + Hotshot-XL)	—	—	—	—	7.68	24.79	42.3	784.3
+ style LoRA	✓	—	—	Partial	6.85	19.52	41.6	550.1
+ style and layout LoRA	✓	✓	—	Partial	6.42	16.81	58.1	480.5
Partial: style + motion	✓	—	✓	Partial	6.59	18.50	41.5	524.9
Partial: layout + motion	—	✓	✓	Partial	7.02	21.80	56.1	509.4
Monolithic LoRA	✓	✓	✓	No	6.35	15.70	49.7	446.4
Ours: style + layout + motion	✓	✓	✓	**Yes**	**6.09**	**14.40**	**62.8**	**413.8**

not interchangeable. Removing the layout adapter keeps BoxAcc@0.5 low, even when style and motion adapters are used. Removing the style adapter degrades image quality, although layout grounding remains relatively better than the baseline. Taken together, these results provide empirical evidence that the proposed architectural separation reduces cross-factor interference compared with monolithic adaptation.

We further compare our method against representative general-purpose video generation baselines, including SVD [11] and CogVideoX [13]. Since these models use different native conditioning formats, we evaluate them under their standard settings while keeping the mining scenario descriptions consistent. CogVideoX is evaluated with the same text prompts. For SVD, which is primarily an image-to-video model, we provide prompt-consistent initial frames and then generate the corresponding video clips. CLIP-score [24] is computed as the average frame-level text-image alignment score over the generated 8-frame clip.

Table 3 highlights that general-purpose video generation models, despite their large-scale pre-training, struggle with the mining-domain gap. They often produce domain-inconsistent machinery appearance, unstable object geometry, or temporally inconsistent frame transitions. In contrast, our factor-isolated framework improves visual fidelity, semantic alignment, and empirical temporal consistency by adapting the visual, spatial, and temporal factors in separate subspaces.

Table 3 Comparison with general-purpose video generation baselines. Bold marks the best metric values. "↑" and "↓" indicate that higher and lower values are better, respectively.

Method	FVD ↓	CLIP-score ↑
SVD	622.3	30.6
CogVideoX	596.1	29.3
Ours	**413.8**	**32.2**

Since FVD and CLIP-score do not directly evaluate the frame-to-frame plausibility of ego-view motion, we further introduce optical-flow-based temporal proxies to quantify motion realism. This evaluation is suitable for our front-view mining videos, where the ego vehicle itself is not visible and some clips may not contain a consistently trackable foreground vehicle. Therefore, instead of relying on object-centered vehicle trajectories, we measure whether the generated visual motion is temporally smooth and statistically close to real mining clips.

For each video clip $V = \{I_t\}_{t=1}^{T}$, we estimate dense optical flow between adjacent frames and compute three metrics. In implementation, the same fixed dense optical-flow estimator is applied to both real and generated videos. Flow magnitudes are normalized by the image diagonal before computing dataset-level statistics. First, temporal warping error measures the average L_1 photometric reconstruction error over valid pixels after warping I_t to I_{t+1} using the estimated optical flow. Second, flow-distribution distance measures the distance between the optical-flow magnitude distribution of generated clips and that of real mining clips, computed as the first-order Wasserstein distance between normalized flow-magnitude histograms. Third, flow-acceleration outlier rate measures the percentage of adjacent frame transitions whose change in average flow magnitude exceeds the empirical 95th percentile observed in real mining clips. This threshold is estimated only from the real reference clips and is fixed for all generated methods.

As shown in Table 4, the real clips provide the empirical reference for ego-view motion statistics. Compared with SVD and CogVideoX, our method produces lower temporal warping error, a flow distribution closer to real mining clips, and fewer acceleration outliers. These results suggest that the proposed temporal adapter improves not only visual continuity but also the empirical plausibility of ego-view motion, including smoother frame-to-frame displacement and fewer abrupt viewpoint changes.

Table 4 Motion realism evaluation based on optical-flow temporal proxies. Real clips are included as an empirical reference. Bold marks the best metric values. "↑" and "↓" indicate that higher and lower values are better, respectively.

Method	Warping error ↓	Flow distance ↓	Acceleration outlier rate (%) ↓
Real clips (reference)	0.047	0.000	5.0
SVD	0.092	0.057	11.8
CogVideoX	0.086	0.049	9.6
Ours	**0.052**	**0.020**	**6.8**

b. Downstream Task Evaluation

To demonstrate practical utility, we evaluate our generated data on two downstream tasks: two-dimensional (2D) object detection and VLM decision-making.

Object detection task. We trained a YOLOv8-s [25] baseline solely on real AutoMine images (real-only) and compared it with a fixed-budget setting that incorporates frames from our generated long-tail videos (real + ours). To avoid conflating the effect of generated data quality with the effect of a larger training set, both settings use the same total number of training images. The detector architecture, preprocessing pipeline, training schedule, and optimization hyperparameters are kept identical across the two settings. Evaluated on a held-out real-world test set that contains no generated images (Table 5), "real + ours" shows tangible mean average precision (mAP) improvements. These results suggest that the gain comes from better coverage of underrepresented rare events and from context-consistent frame-to-frame variations provided by temporally continuous generated sequences.

Table 5 Object detection performance on a real-world test set. Bold marks the best-performing setting.

Training data	mAP@50	mAP@50-95
Real-only (baseline)	58.6	44.7
Real + ours (fixed-budget replacement)	**61.3**	**48.2**

VLM decision-making task. To evaluate whether the generated dynamic scenes are useful beyond low-level perception, we formulate a VLM-based driving decision task for autonomous mining. Given a short video clip and an instruc-

tion prompt, the model is required to identify the dominant hazard and predict an appropriate ego-vehicle action. This task focuses on decision-making from dynamic visual context rather than on generative video quality alone.

We use Qwen2.5-VL-3B [17] as the base VLM. Each input sample consists of an 8-frame clip uniformly sampled from either a real or generated video sequence. The instruction prompt follows a fixed template. Given the mining driving video, identify the main hazard and choose the safest ego-vehicle action. The hazard label space includes adverse weather, abnormal lighting, road-surface hazard, stopped or slow vehicle, abnormal intrusion, and collision risk. The action label space includes continue, slow down, stop, bypass, and yield. We report two metrics: hazard identification accuracy, which measures whether the model identifies the correct hazard category; and action accuracy, which measures whether the predicted driving action matches the ground-truth safety instruction.

As shown in Table 6, we compare three settings: zero-shot inference with the base VLM, instruction tuning with real AutoMine clips only, and instruction tuning with real clips with fixed-budget replacement by generated long-tail sequences. The two instruction-tuning settings used the same number of training clips and the same number of optimization steps. In the augmented setting, generated clips were used to replace a fixed portion of the real-only sampled clips rather than to increase the total training volume.

c. Qualitative Analysis

Fig. 2 showcases the model's ability to render complex static scenes with high fidelity, including adverse weather, abnormal lighting, road-surface hazards, and unusual object configurations. Fig. 3 further illustrates temporally coherent video clips generated after motion adaptation. Compared with the original video backbone, the generated clips exhibit more stable object geometry, smoother frame-to-frame transitions, and more mining-consistent visual motion patterns. These qualitative results support the effectiveness of adapting the temporal layers while preserving the static visual and layout priors learned in sage 1.

Table 6 Downstream VLM decision-making performance on held-out mining video clips. Bold marks the best metric values. "↑" and "↓" indicate that higher and lower values are better, respectively.

Training data	Hazard ID (%) ↑	Action accuracy (%) ↑
Zero-shot base VLM	64.8	52.3
Real-only	72.3	68.0
Real + ours	**86.5**	**83.4**

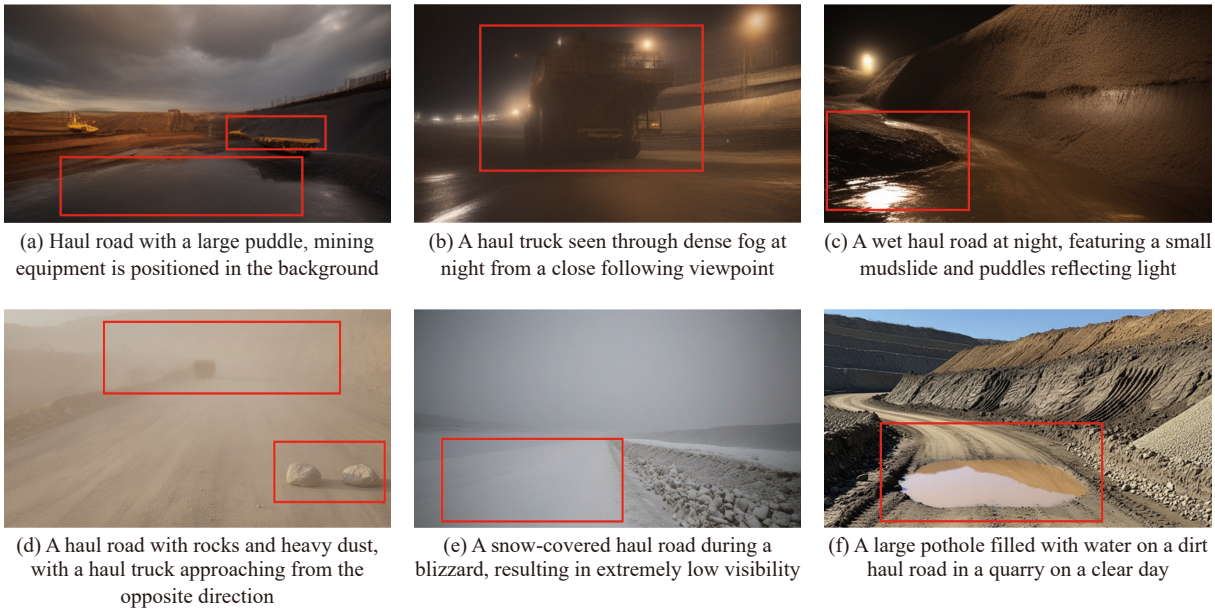
d. Generated Long-Tail Scenario Dataset

To address long-tail data scarcity, we use the proposed pipeline to generate a large-scale dataset containing 5000 distinct video sequences. Each sequence is associated with one or more abnormal attributes from the taxonomy in Table 1. For example, a single sequence may simultaneously involve night-time lighting, dense fog, a wet road surface, and a stopped haul truck. Therefore, the attribute counts shown in Fig. 4 are overlapping multi-label counts rather than mutually exclusive category counts. As a result, the sum of the bars can exceed the total number of generated sequences.

This multi-label design better reflects real mining operations, where safety-critical scenarios often arise from the combination of multiple adverse conditions rather than from a single isolated factor. By providing controllable samples for such combinations, the generated dataset serves as a useful resource for training and evaluating robust autonomous mining systems.

VI. Conclusions

This paper presents a factor-isolated framework for long-tail scenario generation in autonomous mining. By assigning visual style, spatial layout, and temporal motion to separate LoRA subspaces, the method improves controllability and empirically reduces cross-factor interference compared with



(a) Haul road with a large puddle, mining equipment is positioned in the background

(b) A haul truck seen through dense fog at night from a close following viewpoint

(c) A wet haul road at night, featuring a small mudslide and puddles reflecting light

(d) A haul road with rocks and heavy dust, with a haul truck approaching from the opposite direction

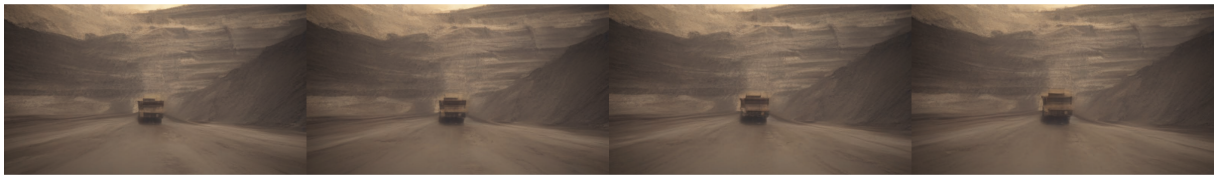
(e) A snow-covered haul road during a blizzard, resulting in extremely low visibility

(f) A large pothole filled with water on a dirt haul road in a quarry on a clear day

Fig. 2 Examples of generated long-tail scenarios using our fine-tuned style and layout LoRA, including combinations of abnormal weather, lighting, and events.

(a) A haul truck approaching from the opposite direction on a dusty road in an open-pit mine during the day

(b) Approaching a blind corner on a haul road



(c) Driving on a steep slope on a haul road, with the haul truck in front appearing to have stopped

Fig. 3 Consecutive frames from generated video clips demonstrating temporal consistency after motion adaptation. Each row represents a distinct scenario, with frames arranged chronologically from left to right to illustrate the progression of events.

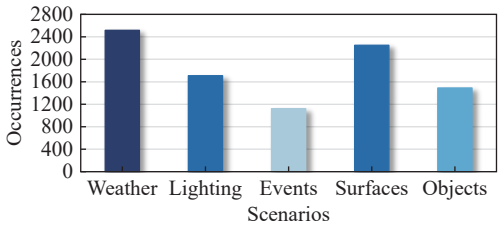


Fig. 4 Statistical distribution of abnormal attributes in the generated dataset. Each generated sequence may contain multiple abnormal attributes.

monolithic adaptation. Experiments with partial and monolithic baselines show improvements in image quality, layout accuracy, and video quality, while downstream evaluations demonstrate gains in object detection and VLM decision-making. The current framework learns empirical motion patterns from data rather than enforcing hard vehicle-dynamics constraints; future work will explore trajectory-conditioned and physics-aware generation.

ACKNOWLEDGMENT

This work was supported by the New Generation Artificial Intelligence National Science and Technology Major Project (No. 2025ZD0124205), the Key Research and Development Program of Xinjiang Uyghur Autonomous Region (No. 2025B01001), and the Key Research and Development Program of Shaanxi Province (No. 2024CY2-GJHX-49).

REFERENCES

[1] S. Ge, F.-Y. Wang, J. Yang, Z. Ding, X. Wang, Y. Li, S. Teng, Z. Liu, Y. Ai, and L. Chen, Making standards for smart mining operations: Intelligent vehicles for autonomous mining transportation, *IEEE Trans. Intell. Veh.*, 2022, 7(3), 413–416.

[2] S. Yang, Y. Huang, L. Li, S. Feng, X. Na, H. Chen, and A. Khajepour, How to guarantee driving safety for autonomous vehicles in a real-world environment: A perspective on self-evolution mechanisms, *IEEE Intell. Transp. Syst. Mag.*, 2024, 16(2), 41–54.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, 10674–10685.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, LoRA: Low-rank adaptation of large language models, in *Proc. 10th International Conference on Learning Representations*, Virtual Event, 2022, 1–17.

[5] Y. Li, Z. Li, S. Teng, Y. Zhang, Y. Zhou, Y. Zhu, D. Cao, B. Tian, Y. Ai, Z. Xuanyuan et al., AutoMine: An unmanned mine dataset, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, 2022, 21276–21285.

[6] J. Mullan, D. Crawbuck, and A. Sastry, Hotshot-XL [Online], https://github.com/hotshotco/hotshot-xl, 25 June 2025.

[7] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, MagicDrive: Street view generation with diverse 3D geometry control, in *Proc. 12th International Conference on Learning Representations*, Vienna, Austria, 2024, 1–20.

[8] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, DriveDreamer-2: LLM-enhanced world models for diverse driving video generation, in *Proc. 39th AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, 2025, 10412–10420.

[9] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang, Panacea: Panoramic and controllable video generation for autonomous driving, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, 6902–6912.

[10] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, Vista: A generalizable driving world model with high fidelity and versatile controllability, in *Proc. 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024, 2906.

[11] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts et al., Stable video diffusion: Scaling latent video diffusion models to large datasets, arXiv preprint arXiv: 2311.15127, 2023.

[12] OpenAI, Video generation models as world simulators [Online], https://openai.com/index/video-generation-models-as-world-simulators/, 3 March 2025.

[13] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng et al., CogVideoX: Text-to-video diffusion models with an expert transformer, in *Proc. 13th International Conference on Learning Representations*, Singapore, Singapore, 2025, 1–30.

[14] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning, in *Proc. 12th International Conference on Learning Representations*, Vienna, Austria, 2024, 1–19.

[15] Z. Liu, J. Wang, Z. Duan, C. Rodriguez-Opazo, and A. van den Hengel, Frame-wise conditioning adaptation for fine-tuning diffusion models in text-to-video prediction, arXiv preprint arXiv: 2503.12953, 2025

[16] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, GLIGEN: Open-set grounded text-to-image generation, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, 22511–22521.

[17] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang et al., Qwen2.5-VL technical report, arXiv preprint arXiv: 2502.13923, 2025.

[18] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., GPT-4 technical report, arXiv preprint arXiv: 2303.08774, 2023.

[19] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen et al., Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, arXiv preprint arXiv: 2507.06261, 2025.

[20] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su et al., Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection, in *Proc. 18th European Conference on Computer Vision*, Milan, Italy, 2024, 38–55.

[21] A. Mittal, R. Soundararajan, and A. C. Bovik, Making a "completely blind" image quality analyzer, *IEEE Signal Process. Lett.*, 2013, 20(3), 209–212.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik, No-reference image quality assessment in the spatial domain, *IEEE Trans. Image Process.*, 2012, 21(12), 4695–4708.

[23] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, Towards accurate generative models of video: A new metric & challenges, arXiv preprint arXiv: 1812.01717, 2018.

[24] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in *Proc. Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, 2021, 7514–7528.

[25] G. Jocher, A. Chaurasia, and J. Qiu, Ultralytics YOLO (Version 8.0.0) [Online], https://github.com/ultralytics/ultralytics, 27 July 2025.

Ruixi Qiao received the BE degree in vehicle engineering from Beijing Institute of Technology, China, in 2020, and the MS degree in automation engineering from RWTH Aachen University, Germany, in 2022. From 2022 to 2024, he worked as a software engineer at Huawei Technologies Co., Ltd. He is currently pursuing the PhD degree in pattern recognition and intelligent systems at Institute of Automation, Chinese Academy of Sciences, China, and University of Chinese Academy of Sciences, China. His research interests include reinforcement learning, large language models, and autonomous driving systems.

Chenglong Ye received the BS degree in transportation engineering from Sun Yat-sen University, Shenzhen, China, in 2024. He is currently pursuing the PhD degree in pattern recognition and intelligent systems at State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include autonomous driving and trajectory planning.

Jie Cheng received the BE degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2022. He is currently pursuing the PhD degree in management and control of complex systems at Institute of Automation, Chinese Academy of Sciences, Beijing, China, and University of Chinese Academy of Sciences, Beijing, China. His research interests include reinforcement learning, offline reinforcement learning, and large language model.

Peizhe Gong received the BS degree in artificial intelligence from University of Chinese Academy of Sciences, Beijing, China, in 2025. He is currently pursuing the PhD degree in computer application technology at Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include scenario generation for autonomous driving.

Xingyuan Dai received the PhD degree in control theory and control engineering from University of Chinese Academy of Sciences, Beijing, China, in 2022. He is currently an assistant professor at State Key Laboratory of Multimodal Artificial Intelligence Systems, Chinese Academy of Sciences, China. His research interests include intelligent transportation systems, intelligent vehicles, and deep reinforcement learning.

Xiaoyan Gong received the PhD degree in control theory and control engineering from Chinese Academy of Sciences, China, in 2003. She is currently a senior engineer at Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her research interests include traffic data analysis, traffic modeling and prediction, and parallel traffic management and control systems.

Yisheng Lv received the PhD degree in control theory and control engineering from Chinese Academy of Sciences, China, in 2010. He is currently a professor at State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China. His research interests include traffic data analysis, autonomous driving, and parallel traffic management and control systems.