FOCALBRIDGE: BRIDGING FOCAL PHYSICS AND GEOMETRIC DRIFT FOR CONTINUOUS TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

While camera-aware text-to-image generation has attracted increasing attention, geometric drift remains a critical challenge in continuous focal-length generation, as small geometric deviations can accumulate across focal transitions and lead to composition distortion and scene inconsistency. Existing approaches typically encode focal length as an implicit camera condition, without explicitly regulating the geometric relationship between focal length, field of view, and scene structure during generation. To address this issue, we propose **FocalBridge**, a physics-guided framework that progressively recovers the focal trajectory, corrects focal-dependent geometry, and preserves high-frequency scene appearance throughout the correction process. Specifically, **Recovery-LF** stabilizes free-running generation by adapting the focal-control pathway to self-generated model states, while a **Geometry-Control Decoder Adapter** introduces explicit physical guidance through bounded geometry correction in the frozen VAE decoder. A **High-Frequency Appearance Preservation** objective further reduces local appearance degradation during correction. Extensive experiments show that FocalBridge substantially improves focal geometry and cross-focal scene consistency over representative baselines, with strong generalization across different scene distributions.

1 INTRODUCTION

Text-to-image generation has attracted increasing attention for its ability to synthesize diverse and realistic visual content from natural language descriptions. Diffusion-based generative models have played a central role in this progress and have substantially improved image quality and semantic fidelity (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022). Despite these advances, fine-grained photographic control remains challenging, particularly when camera parameters are expected to vary continuously while the underlying scene remains stable. For example, changing focal length should produce predictable variations in field of view and image scale while preserving the same scene content. However, existing diffusion-based generative models often entangle focal variation with semantic and compositional changes, leading to inconsistent object scale, layout shifts, and scene variation across focal settings. Although recent studies have improved camera-aware generation (Yuan et al., 2025; Fang et al., 2024; Sun et al., 2025), stable continuous focal control under free-running generation remains insufficiently explored. This gap calls for a generation mechanism that adapts to self-generated states while explicitly regulating focal-dependent geometry.

Recent controllable generation methods provide effective mechanisms for injecting external conditions into pretrained generative models, but they still exhibit important limitations for continuous focal control. On the one hand, spatial-control methods such as ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) incorporate structural conditions including edges, depth, segmentation maps, and poses, enabling explicit control over spatial content. However, these conditions mainly describe external scene structure and do not directly model the physical dependency between camera parameters and image formation. On the other hand, focal-length control requires the generated scene to follow a continuous geometric response determined by focal length and field of view. Existing camera-conditioned methods typically encode focal length as an additional learned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

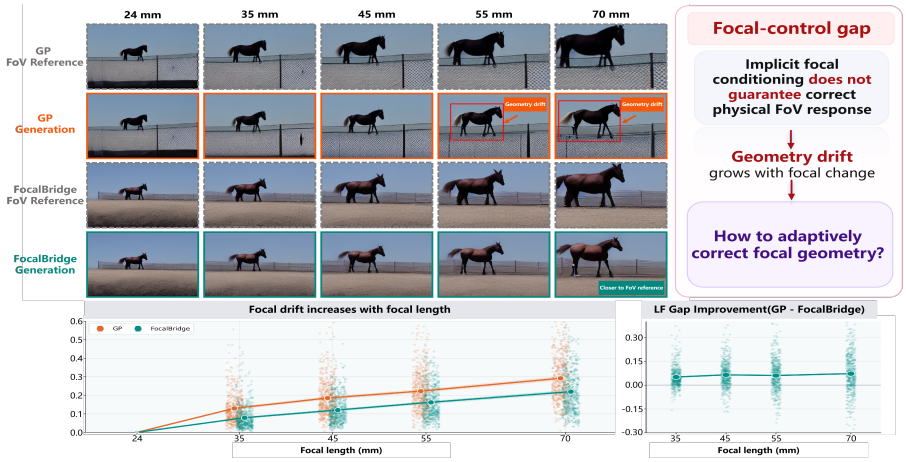


Figure 1: Qualitative and quantitative comparison of focal control. For each method, the analytic FoV reference is derived from its own 24 mm anchor. GP increasingly deviates from the expected focal geometry as focal length grows, whereas FocalBridge more closely follows the corresponding analytic reference. Bottom: LF Gap across focal lengths and the paired improvement over GP.

condition, without explicitly enforcing the corresponding physical geometry during free-running generation. Consequently, deviations in self-generated intermediate states may accumulate along the generation trajectory and lead to geometric drift and scene inconsistency.

Camera-aware text-to-image generation has recently begun to address this problem more directly. Generative Photography (GP) (Yuan et al., 2025) introduces camera conditioning and focal-dependent data construction for scene-consistent generation, while Camera Settings as Tokens (Fang et al., 2024) represents photographic parameters as explicit generative tokens. Although these methods incorporate focal-dependent priors through data construction or explicit conditioning, such priors are not explicitly reused as a generation-time reference to correct geometric deviations arising from self-generated states.

To tackle this limitation, we propose **FocalBridge**, a physics-guided framework for stable continuous focal-length generation. First, **Low-Frequency Focal Trajectory Recovery (Recovery-LF)** adapts the focal-control pathway to self-generated diffusion states, reducing the mismatch between training-time supervision and free-running generation. Second, **Physics-Guided Geometry Correction** converts the focal-dependent prior into an explicit generation-time physical reference and injects bounded geometry corrections into the frozen Variational Autoencoder (VAE) decoder. Finally, **High-Frequency Appearance Preservation** constrains the corrected output to retain target-specific local structures during geometry correction. Extensive experiments validate the effectiveness and generalization of the proposed framework. Our contributions are summarized as follows:

- A **Low-Frequency Focal Trajectory Recovery (Recovery-LF)** strategy is developed to address geometric drift under free-running generation. It adapts the focal-control pathway to self-generated diffusion states and improves the stability of focal-dependent low-frequency structure.
- A **Physics-Guided Geometry Correction** module is designed to address inaccurate focal-dependent geometry. It transforms the focal-dependent prior into an explicit generation-time physical reference and injects a bounded geometry residual into the frozen VAE decoder for direct spatial correction.
- A **High-Frequency Appearance Preservation** objective is introduced to reduce local appearance degradation caused by geometry correction. It preserves target-specific high-frequency structures while retaining the geometric benefit of physical correction.
- Extensive experiments on Formal80, COCO-val500, and different focal ranges demonstrate substantial improvements in focal geometry, particularly for medium-to-large focal changes, together with improved cross-focal scene consistency.

2 Related Work

Diffusion-Based Generation and Control. Diffusion models have become a dominant paradigm for high-quality image and video generation (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Peebles & Xie, 2023; Blattmann et al., 2023; Guo et al., 2024). Existing controllable generation methods introduce external conditions into pretrained diffusion models, including edges, depth, poses, layouts, segmentation masks, reference images, and editing instructions (Meng et al., 2022; Brooks et al., 2023; Zhang et al., 2023; Mou et al., 2024; Li et al., 2023; Ye et al., 2023; Mo et al., 2024; Luo et al., 2024; Wang et al., 2024a). These conditions can be incorporated through control branches, lightweight adapters, attention modulation, or sampling guidance, enabling flexible control over generated content and spatial structure. Related approaches have also been extended to video generation, where motion representations and camera trajectories regulate dynamic content and camera motion (He et al., 2025; Wang et al., 2024b; Xi et al., 2025). These studies establish a general foundation for conditional visual generation, but mainly focus on injecting control signals rather than explicitly regulating camera-dependent geometry during generation.

Generative Photography and Focal Control. Recent camera-aware generation methods introduce photographic or camera parameters into generative models for finer control over image formation. GP (Yuan et al., 2025) incorporates focal-dependent priors through camera conditioning and data construction, Camera Settings as Tokens (Fang et al., 2024) represents photographic settings as explicit conditioning tokens, and PreciseCam (Bernal-Berdun et al., 2025) controls intrinsic and extrinsic camera parameters. Viewpoint- and multi-view-conditioned diffusion further explores camera control and geometric consistency (Liu et al., 2023; Kant et al., 2024; Yang et al., 2024b; Gu et al., 2024; Yang et al., 2024a; Cao et al., 2025). For video generation, CameraCtrl (He et al., 2025), MotionCtrl (Wang et al., 2024b), and CineCtrl (Sun et al., 2025) investigate camera trajectory, viewpoint, and cinematic control. Despite these advances, continuous focal geometry is still largely governed by learned conditioning rather than an explicit generation-time physical reference. FocalBridge instead regulates focal-dependent geometry during generation while preserving scene appearance.

3 Preliminary

Let $\mathbf{x}_s \in \mathcal{X}$ denote an image generated from a text prompt $\mathbf{y}$ under focal length f_s, where $s \in [0, 1]$ indexes a continuous focal trajectory. Given the same text prompt and initial noise $\boldsymbol{\epsilon}$, the goal of focal-controlled generation is to produce a smoothly varying family of conditional distributions $p_\theta(\mathbf{x}_s \mid \mathbf{y}, f_s)$ such that focal-dependent geometry follows the prescribed optical variation while scene appearance remains stable across different focal settings. We denote by $\mathcal{T}_{s_1 \to s_2}$ the physical field-of-view transformation from focal state s_1 to s_2.

Definition 1 (Continuous Focal Deviation (CFD))**.** *During free-running focal generation, deviations in the reverse diffusion trajectory may lead to inaccurate focal geometry and undesired appearance variation. To isolate focal-dependent variation from sampling randomness, different focal states are generated using the same initial noise. We define the Continuous Focal Deviation (CFD), denoted by $\Delta(s_1, s_2)$, between two focal states $s_1, s_2 \in [0, 1]$ as*

$$\begin{aligned}\Delta(s_1, s_2) \triangleq\ & \underbrace{\lambda_{\text{phy}}\, \mathbb{E}_{\boldsymbol{\epsilon}} \left[\mathcal{D}_{\text{phy}} \left(\mathbf{x}_{s_1}, \mathbf{x}_{s_2}; f_{s_1}, f_{s_2}\right)\right]}_{\textit{Focal-Geometry Deviation}} \\ & + \underbrace{\lambda_{\text{app}}\, \mathbb{E}_{\boldsymbol{\epsilon}} \left[\mathcal{D}_{\text{app}} \left(\mathcal{H}\left(\mathcal{T}_{s_1 \to s_2}(\mathbf{x}_{s_1})\right), \mathcal{H}(\mathbf{x}_{s_2})\right)\right]}_{\textit{Appearance Deviation}}.\end{aligned} \tag{1}$$

where $\mathcal{D}_{\text{phy}}$ measures whether the generated transition follows the prescribed focal-dependent field-of-view and scale variation, while $\mathcal{D}_{\text{app}}$ measures local appearance deviation after compensating for the expected focal transformation. Here, $\mathcal{H}(\cdot)$ extracts high-frequency image structures such as local edges and texture variations. The coefficients $\lambda_{\text{phy}}, \lambda_{\text{app}} > 0$ balance focal-geometry regulation and appearance preservation, respectively.

Standard focal-conditioned diffusion does not explicitly guarantee that the generated sequence remains physically stable throughout free-running reverse diffusion. For each focal state f_s, the de-

noising process evolves through $\mathbf{z}_T \to \cdots \to \mathbf{z}_0$, where each state depends on previous model predictions. Small deviations at intermediate timesteps may therefore propagate to subsequent states and accumulate into geometric drift and local appearance degradation in the generated sequence. For a finite focal interval $\delta > 0$, we define the integrated local deviation as

$$\mathcal{J}_{\mathrm{CFD}}(\delta) = \int_0^{1-\delta} \Delta(s, s+\delta)\, ds. \tag{2}$$

We use the term *geometric drift* to describe the progressive accumulation of focal-geometry deviation during free-running generation. Specifically, it refers to the mismatch between the generated focal response and the physically expected field-of-view, image-scale, and spatial-layout variation as focal length changes.

This formulation gives rise to the following three challenges:

Challenge 1 (Free-Running Geometric Drift)**.** *During reverse diffusion,*

$$\mathbf{z}_{t-1} \sim p_\theta\left(\cdot \mid \mathbf{z}_t, \mathbf{y}, f_s\right), \tag{3}$$

where each denoising state is generated from the model's previous prediction. Consequently, small focal-dependent deviations may move the generation trajectory away from the states observed during training and propagate through subsequent denoising steps. Preventing the accumulation of geometric drift under such self-generated states is therefore a fundamental challenge for continuous focal generation.

Challenge 2 (Physical Geometry Regulation)**.** *Focal-length variation follows an explicit optical relationship among focal length, field of view, image scale, and visible spatial extent. However, focal conditioning alone does not guarantee that the generated image follows this physical relationship accurately. The core challenge is to introduce explicit focal-dependent geometric information into the generation process while preserving the original semantic and rendering capabilities of the pretrained model.*

Challenge 3 (High-Frequency Appearance Preservation)**.** *Explicit geometry correction modifies the spatial structure of the generated image and may simultaneously alter target-specific local appearance, including edges, textures, and fine structures. Such changes are not required by focal variation itself and may reduce the visual consistency of the generated focal sequence. The key challenge is therefore to preserve the high-frequency characteristics of the original target view while correcting its low-frequency focal geometry.*

4 METHODOLOGY

We present **FocalBridge** to address three key challenges in continuous focal-length control. It integrates **Recovery-LF**, **Physics-Guided Geometry Correction**, and **High-Frequency Appearance Preservation** into a unified framework, progressively stabilizing the free-running focal trajectory, correcting focal-dependent geometry with explicit physical guidance, and preserving target-specific high-frequency appearance during geometry correction.

RECOVERY-LF

To address free-running geometric drift (**Challenge 1**), **Recovery-LF** adapts the focal-control pathway to self-generated intermediate diffusion states, reducing the mismatch between teacher-forced training and free-running generation.

1. Self-Generated Focal State Construction. Given a paired clean latent $\mathbf{z}_{i,0}$ at focal length f_i, we first construct a noisy state at timestep $t + \Delta t$ and let the current model perform one deterministic DDIM transition to timestep t:

$$\tilde{\mathbf{z}}_{i,t} = \mathrm{DDIM}_\theta\left(\mathbf{z}_{i,t+\Delta t}, t{+}\Delta t \to t; \mathbf{y}, f_i\right). \tag{4}$$

Unlike a directly sampled teacher-forced state, $\tilde{\mathbf{z}}_{i,t}$ contains deviations introduced by the model itself. Using its paired clean latent, we recover the corresponding supervision target as $\boldsymbol{\epsilon}^*_{i,t} = \frac{\tilde{\mathbf{z}}_{i,t} - \sqrt{\bar{\alpha}_t}\mathbf{z}_{i,0}}{\sqrt{1-\bar{\alpha}_t}}$, providing a denoising target for the encountered free-running state.

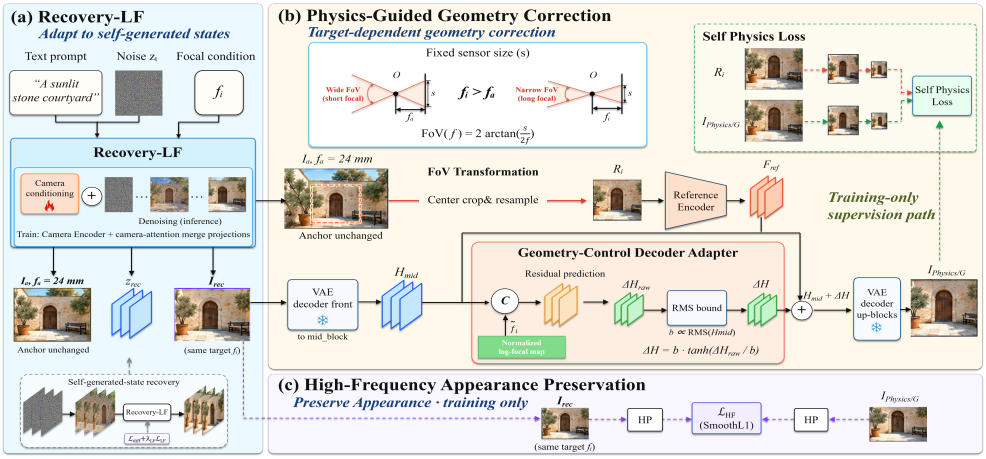


Figure 2: Overview of the proposed FocalBridge framework.

2. Low-Frequency Focal Recovery. Given $\tilde{\mathbf{z}}_{i,t}$, the model predicts $\hat{\boldsymbol{\epsilon}}_{i,t}$ and reconstructs the clean latent as $\hat{\mathbf{z}}_{i,0} = \frac{\tilde{\mathbf{z}}_{i,t} - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_{i,t}}{\sqrt{\bar{\alpha}_t}}$. Recovery-LF jointly preserves the diffusion objective and introduces a low-frequency recovery constraint:

$$\mathcal{L}_{\text{Recovery}} = \underbrace{\left\|\hat{\boldsymbol{\epsilon}}_{i,t} - \boldsymbol{\epsilon}^*_{i,t}\right\|_2^2}_{\mathcal{L}_{\text{diff}}} + \lambda_{\text{rec}} \underbrace{\mathcal{D}_{\text{LF}}\left(\hat{\mathbf{z}}_{i,0}, \mathbf{z}_{i,0}\right)}_{\mathcal{L}_{\text{rec}}}. \tag{5}$$

Here, $\mathcal{D}_{\text{LF}}$ compares the channel-averaged, pooled latent intensity and its spatial gradients using cosine distance. The recovery term is applied only to the non-anchor focal views, encouraging recovery of focal-dependent low-frequency structure from self-generated states.

3. Focal-Control Pathway Adaptation. Recovery-LF fine-tunes the complete camera encoder together with the native `qkv_merge` projections in all temporal camera-conditioned attention processors, while freezing the remaining UNet, VAE, and text encoder. At inference time, the recovered camera-control checkpoint replaces the corresponding GP control weights to jointly generate the multi-focal sequence. Recovery-LF itself does not hard-bypass the anchor view; once generated, however, its widest-angle output is adopted as the fixed scene anchor and preserved unchanged by the subsequent geometry-correction stage.

Physics-Guided Geometry Correction

To correct focal-dependent geometry (**Challenge 2**), **Physics-Guided Geometry Correction** derives a physical reference from the wide-angle anchor and injects bounded residuals into the frozen VAE decoder.

1. Focal-Dependent Physical Reference. Let I_a^{rec} denote the anchor generated by Recovery-LF with focal length f_a. Following GP's official angular-FoV crop convention, we compute the target-dependent crop ratio as

$$s_i = \min_{d \in \{x,y\}} \frac{\arctan\left(\frac{h_d}{f_i}\right)}{\arctan\left(\frac{h_d}{f_a}\right)}, \tag{6}$$

where h_d denotes the half sensor size along dimension d. Physics/G follows this crop-ratio convention using differentiable affine-grid sampling, yielding a target-focal RGB reference R_i. A lightweight reference encoder further maps R_i into F_i^{ref}, transforming the focal prior from an offline data-construction rule into an explicit generation-time geometry reference.

2. Geometry-Control Decoder Adapter. Given the Recovery-LF latent z_i^{rec}, the frozen VAE decoder produces a mid-block representation H_i^{mid}. The target focal length is encoded as $\tilde{f}_i = \frac{\log(f_i/f_a)}{\log(f_{\max}/f_a)}$ and spatially broadcast to match the decoder feature. We then jointly fuse the current decoder state,

physical reference, and focal condition:

$$\Delta H_i^{\text{raw}} = \mathcal{G}\left(\left[H_i^{\text{mid}}; F_i^{\text{ref}}; \tilde{f}_i \mathbf{1}\right]\right), \tag{7}$$

where $\mathcal{G}(\cdot)$ denotes the proposed Geometry-Control Decoder Adapter. Its final convolution is zero-initialized, making the initial residual exactly zero. This allows target-specific geometry correction to depend jointly on the current decoder state, physical reference, and focal condition.

3. Bounded Geometry Correction. To constrain the correction magnitude relative to the current decoder representation, we set $b_i = \lambda_b \max(\text{RMS}(H_i^{\text{mid}}), \varepsilon)$ and inject the bounded residual immediately after the complete VAE mid-block:

$$I_i^{\text{phy}} = D_{\text{up}}\left(H_i^{\text{mid}} + b_i \tanh\left(\frac{\Delta H_i^{\text{raw}}}{b_i}\right)\right), \tag{8}$$

where D_{up} denotes the remaining frozen VAE decoder. Physics/G performs focal correction directly in decoder feature space. The geometry-control path is bypassed for the anchor view.

High-Frequency Appearance Preservation

To prevent local appearance degradation during geometry correction (**Challenge 3**), we constrain the corrected output to preserve the high-frequency structures of the corresponding Recovery-LF output.

1. High-Frequency Appearance Reference. Given the Recovery-LF output I_i^{rec}, we extract its high-frequency component using a local low-pass residual:

$$\mathcal{H}(I) = I - \text{AvgPool}_{k_h}(I), \tag{9}$$

where k_h denotes the low-pass kernel size. The resulting high-frequency representation mainly captures local edges, textures, and fine structures already established by Recovery-LF, and is used as the appearance reference during geometry correction.

2. Appearance Preservation Objective. Given the geometry-corrected output $I_i^{G,\text{self}}$, we constrain its high-frequency component to remain close to that of the corresponding Recovery-LF output:

$$\mathcal{L}_{\text{HF}} = \mathcal{L}_{\text{s1}}\left(\mathcal{H}(I_i^{G,\text{self}}), \mathcal{H}(I_i^{\text{rec}})\right), \tag{10}$$

where $\mathcal{L}_{\text{s1}}$ denotes the Smooth L1 loss. This objective reduces unnecessary changes in target-specific local appearance introduced by decoder-space geometry correction.

3. Joint Geometry-Appearance Optimization. The high-frequency preservation objective is jointly optimized with the self-supervised geometry constraint:

$$\mathcal{L}_{\text{self}} = \mathcal{L}_{\text{geo}} + \lambda_{\text{HF}} \mathcal{L}_{\text{HF}}, \tag{11}$$

allowing Physics/G to correct low-frequency focal geometry while retaining the high-frequency appearance inherited from Recovery-LF. Since this objective is used only during training, it introduces no additional network parameters or inference-time processing.

4.1 Training Objective

Recovery-LF and Physics/G are optimized sequentially. Recovery-LF is first trained with the objective in Eq. equation 5. We then freeze the recovered focal generator and optimize the Geometry-Control Decoder Adapter using

$$\mathcal{L}_G = \mathcal{L}_{\text{self}} + \lambda_{\text{pair}} \mathcal{L}_{\text{pair}} + \lambda_{\text{res}} \mathcal{L}_{\text{res}}, \tag{12}$$

where $\mathcal{L}_{\text{self}} = \mathcal{L}_{\text{geo}}(I^{G,\text{self}}, R^{\text{self}}) + \lambda_{\text{HF}} \mathcal{L}_{\text{HF}}$ enforces focal geometry on free-running Recovery-LF outputs while preserving their high-frequency appearance, $\mathcal{L}_{\text{pair}} = \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{geo}}(I^{G,\text{pair}}, I^{\text{GT}})$ provides paired image and geometry supervision, and $\mathcal{L}_{\text{res}} = \text{mean}[(\Delta H)^2]$ regularizes the decoder residual. Here, $\mathcal{L}_{\text{HF}} = \mathcal{L}_{\text{s1}}\left(\mathcal{H}(I^{G,\text{self}}), \mathcal{H}(I^{\text{rec}})\right)$ and $\mathcal{L}_{\text{RGB}} = \mathcal{L}_{\text{s1}}\left(I^{G,\text{pair}}, I^{\text{GT}}\right)$.

The shared multi-scale geometry loss is defined as

$$\mathcal{L}_{\text{geo}}(P,T) = \frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\left[\mathcal{L}_{\text{s1}}(P_k,T_k) + \lambda_{\nabla}\sum_{d\in\{x,y\}}\mathcal{L}_{\text{s1}}\left(\partial_d P_k, \partial_d T_k\right)\right], \tag{13}$$

where P_k and T_k denote average-pooled predictions and targets at scale k. Physics/G is optimized only on the non-anchor focal views, while the VAE and Recovery-LF generator remain frozen.

5 Theoretical Analysis

Theorem 1 (Bounded Geometry Correction)**.** *Let H_i^{mid} denote the frozen VAE mid-block representation and let*

$$\Delta H_i = b_i \tanh\left(\frac{\Delta H_i^{\text{raw}}}{b_i}\right), \qquad b_i = \lambda_b \max\left(\text{RMS}(H_i^{\text{mid}}), \varepsilon\right).$$

If the remaining frozen decoder D_{up} is L_D-Lipschitz, with $I_i^{\text{base}} = D_{\text{up}}(H_i^{\text{mid}})$, then

$$\|\Delta H_i\|_\infty \le b_i, \qquad \left\|I_i^{\text{phy}} - I_i^{\text{base}}\right\|_2 \le L_D \|\Delta H_i\|_2 \le L_D\sqrt{CHW}\, b_i. \tag{14}$$

Remark 1. *The bounded residual limits the magnitude of decoder-space geometry correction relative to the current VAE representation, while the high-frequency preservation objective constrains local appearance changes during training.*

6 Experiments

We evaluate FocalBridge from four complementary perspectives: *(1)* comparison with representative text-to-image and camera-aware generation methods, *(2)* component-wise analysis of Recovery-LF, Physics/G, and High-Frequency Appearance Preservation, *(3)* cross-dataset and focal-range robustness, and *(4)* qualitative and efficiency analysis.

6.1 Overall Performance

We introduce LF Gap to evaluate absolute focal-geometry accuracy and follow GP (Yuan et al., 2025) in reporting CorrCoef for focal-trajectory consistency. FOV-aligned DINOv2 and CLIP measure cross-focal scene preservation and text-image alignment, respectively. Table 1 also reports the component ablation of FocalBridge under the same evaluation protocol.

FocalBridge achieves the lowest LF Gap of 0.1222 among the compared generative methods, reducing focal-geometry error by 37.1% relative to GP and by 33.1% relative to CameraCtrl-FT-Focal. It also obtains the highest CorrCoef of 0.9108 and DINOv2 score of 0.6349, while maintaining comparable text-image alignment. These results demonstrate improved focal accuracy and cross-focal scene consistency. The physical simulator achieves a near-zero LF Gap of 0.0013, providing an analytic reference for physically consistent focal transformations.

6.2 Experimental Settings

We evaluate FocalBridge on the held-out Formal80 benchmark and a fixed 500-scene subset of COCO-val. The main comparison uses five focal settings $\{24, 35, 45, 55, 70\}$ mm, where 24 mm serves as the anchor view. All compared methods use identical prompts, focal settings, random seeds, and output resolutions within each experiment.

We compare against Stable Diffusion 3 (Esser et al., 2024), FLUX.1-dev (Labs, 2024), Camera Settings as Tokens (Fang et al., 2024), CameraCtrl (He et al., 2025), and GP (Yuan et al., 2025). CameraCtrl is fine-tuned for focal-length control following the GP comparison setting and is denoted as CameraCtrl-FT-Focal throughout the experiments. For evaluation, we introduce **LF Gap (Low-Frequency Geometry Gap)** to measure whether the generated focal-dependent field-of-view and scale changes agree with the analytic camera transformation derived from each method's own

Table 1: Quantitative comparison and component ablation on Formal80. Best generative results are in bold and second-best results are underlined.

Method	Control	Focal Geometry		Scene Preservation	
		LF Gap↓	CorrCoef↑	DINOv2↑	CLIP↑
Physical Simulator	Physical Ref.	0.0013	0.9999	–	–
Stable Diffusion 3 (2024)	Text	0.8516	−0.0342	0.3588	**0.3198**
FLUX.1-dev (2024)	Text	0.8013	0.0827	0.4069	<u>0.3163</u>
Camera Settings as Tokens (2024)	Camera	0.8163	0.1062	0.3836	0.3120
CameraCtrl-FT-Focal (2025)	Camera	<u>0.1828</u>	<u>0.8999</u>	<u>0.6224</u>	0.2898
GP (2025)	Multi-Focal	0.1944	0.8757	0.6209	0.2853
FocalBridge (Ours)	**Multi-Focal**	**0.1222**	**0.9108**	**0.6349**	0.2867
Component Ablation					
GP baseline	–	0.1944	0.8757	0.6209	0.2853
GP + Physics/G	–	0.1648	0.9030	0.6245	0.2852
GP + Recovery-LF	–	0.1474	0.8874	0.6308	0.2872
Recovery-LF + Physics/G	–	0.1231	0.9066	0.6342	0.2865
FocalBridge	–	**0.1222**	**0.9108**	**0.6349**	0.2867

Table 2: Cross-dataset comparison on 500 COCO-val scenes. The best result is shown in bold and the second-best is underlined.

Method	LF Gap↓	CorrCoef↑	DINOv2↑	CLIP↑
Stable Diffusion 3	0.8188	−0.0230	0.3690	<u>0.3059</u>
FLUX.1-dev	0.7857	−0.0083	0.4071	**0.3075**
Camera Settings as Tokens	0.8259	0.0229	0.3822	0.2995
CameraCtrl-FT-Focal	0.2474	0.8323	0.6108	0.2789
GP	<u>0.2076</u>	<u>0.8746</u>	<u>0.6216</u>	0.2788
FocalBridge	**0.1457**	**0.8885**	**0.6334**	0.2797

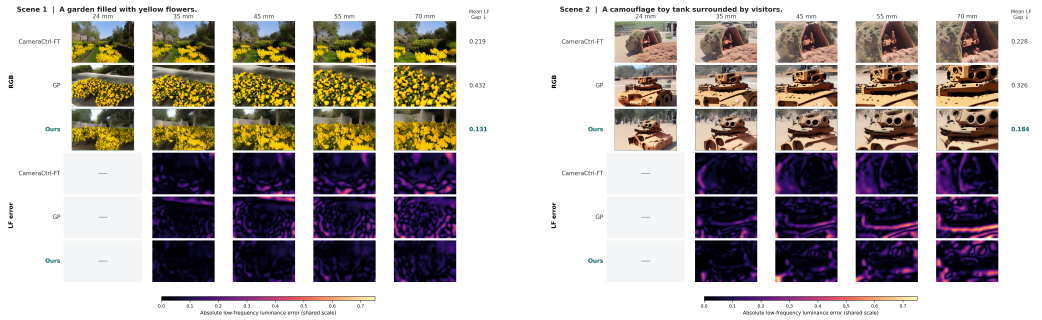
anchor view. For an input focal sequence, LF Gap is computed on all non-anchor views and then averaged at the scene and dataset levels. Lower values indicate more accurate focal geometry, with an ideal value of zero. Following GP (Yuan et al., 2025), we additionally report CorrCoef for focal-trajectory consistency using the analytic focal trajectory derived from each method's own generated 24 mm anchor. FOV-aligned DINOv2 and CLIP evaluate cross-focal scene preservation and text-image alignment, respectively. Detailed metric definitions and evaluation protocols are provided in Appendix B.

Recovery-LF fine-tunes only the focal-control pathway while keeping the remaining generative backbone frozen. Physics/G is subsequently optimized with the recovered focal generator fixed, while High-Frequency Appearance Preservation is used only during training and introduces no additional inference-time parameters or processing. Additional dataset and implementation details are provided in Appendix A.

6.3 Experiments Discussion

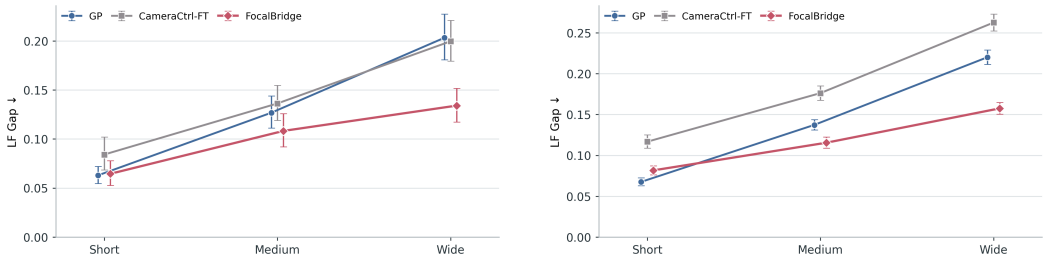
Component analysis. As shown in Table 1, both Recovery-LF and Physics/G independently improve focal control. Recovery-LF mainly improves focal stability under self-generated states, while Physics/G provides explicit generation-time geometry correction. Combining the two components yields further gains, demonstrating their complementarity. High-Frequency Appearance Preservation provides an additional consistent improvement in focal geometry and scene preservation while keeping text-image alignment nearly unchanged. Detailed component results are provided in Appendix C.

Qualitative comparison. Figure 3 presents representative focal sequences and their corresponding low-frequency geometry-error maps. Compared with GP and CameraCtrl-FT-Focal, FocalBridge



(a) Scenario 1 (b) Scenario 2

Figure 3: Qualitative comparison on representative Formal80 scenes. RGB sequences and corresponding LF error maps are shown across focal settings.



(a) Formal80 (b) COCO-val500

Figure 4: Focal-range robustness on Formal80 and COCO-val500 across short, medium, and wide focal spans.

more closely follows the expected focal transformation, particularly at larger focal lengths, while maintaining the overall scene structure.

Cross-dataset generalization. As shown in Table 2, FocalBridge achieves the lowest LF Gap and the highest CorrCoef and DINOv2 among the compared generative methods on COCO-val500. The consistent improvement over GP indicates that the proposed focal-control mechanism generalizes beyond the Formal80 benchmark.

Focal-range robustness. As shown in Figure 4, FocalBridge provides increasingly clear geometry improvements as the focal range becomes larger. Detailed results for short, medium, and wide focal trajectories, together with scene-level bootstrap confidence intervals, are reported in Appendix D.

Efficiency. FocalBridge introduces no additional diffusion evaluations and incurs only a modest inference overhead over GP. Detailed runtime analysis is provided in Appendix E.

7 Conclusion

We presented **FocalBridge**, a physics-guided framework for continuous focal-length control in text-to-image generation. FocalBridge combines Recovery-LF for stabilizing focal responses under self-generated states, Physics-Guided Geometry Correction for explicit generation-time focal regulation, and High-Frequency Appearance Preservation for reducing local appearance degradation during correction. Experiments on Formal80 and COCO-val500 demonstrate substantial improvements in focal geometry and cross-focal scene consistency over representative text-to-image and camera-aware generation methods. A current limitation is that the physical reference is constructed from a fixed wide-angle anchor using an analytic field-of-view transformation, which cannot fully model depth-dependent parallax, newly revealed regions, or complex occlusion changes. Future work may incorporate richer scene geometry while preserving the lightweight frozen-backbone design.

REFERENCES

Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. PreciseCam: Precise camera control for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2724–2733, 2025.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6045–6056, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 12606–12633, 2024.

I-Sheng Fang, Yue-Hua Han, and Jun-Cheng Chen. Camera settings as tokens: Modeling photography on latent diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 62:1–62:11, 2024.

Yuming Gu, Hongyi Xu, You Xie, Guoxian Song, Yichun Shi, Di Chang, Jing Yang, and Linjie Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10456–10465, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multiview diffusers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10026–10038, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8217–8227, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7465–7475, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Huiqiang Sun, Liao Shen, Zhan Peng, Kun Wang, Size Wu, Yuhang Zang, Tianqi Liu, Zihao Huang, Xingyu Zeng, Zhiguo Cao, Wei Li, and Chen Change Loy. CineCtrl: Generative photographic control for scene-consistent video cinematic editing. *arXiv preprint arXiv:2511.12921*, 2025.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6232–6242, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 114:1–114:11, 2024b.

Dianbing Xi, Jiepeng Wang, Yuanzhi Liang, Xi Qiu, Jialun Liu, Hao Pan, Yuchi Huo, Rui Wang, Haibin Huang, Chi Zhang, and Xuelong Li. Ctrlvdiff: Controllable video generation via unified multimodal video diffusion, 2025.

Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7079–7088, 2024a.

Xianghui Yang, Yan Zuo, Sameera Ramasinghe, Loris Bazzani, Gil Avraham, and Anton van den Hengel. Viewfusion: Towards multi-view consistency via interpolated denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9870–9880, 2024b.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley H. Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7920–7930, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

A Additional Experimental Details

Formal80 benchmark. We use Formal80 as the primary controlled benchmark. It contains 80 held-out prompts constructed from the GP `camera_bokehK` scene pool, excluding scenes used during method development, calibration, or coupling selection. For the main comparison, each prompt is generated at a resolution of 384×256 under five focal settings $\{24, 35, 45, 55, 70\}$ mm, where 24 mm serves as the anchor view. All compared methods use identical prompts, focal settings, random seeds, and output resolutions.

COCO-val500. To evaluate generalization beyond the GP scene distribution, we additionally use a fixed 500-scene subset of COCO-val. The same five-view focal sequence $\{24, 35, 45, 55, 70\}$ mm is used for the main comparison. Within each experiment, all compared methods use identical prompts, focal settings, random seeds, and output resolutions, enabling direct comparison under matched sampling conditions.

Baselines. We compare FocalBridge against Stable Diffusion 3, FLUX.1-dev, Camera Settings as Tokens, CameraCtrl, and GP. For GP, we use the official implementation and released focal checkpoint without retraining. CameraCtrl is fine-tuned for focal-length control following the GP comparison setting and is denoted as CameraCtrl-FT-Focal throughout the experiments.

Evaluation metrics. We introduce **LF Gap (Low-Frequency Geometry Gap)** to evaluate whether the generated focal-dependent field-of-view and scale changes agree with the analytic camera transformation derived from each method's own anchor view. For the main Formal80 and COCO-val500 comparisons, LF Gap is computed over the four non-anchor focal views $\{35, 45, 55, 70\}$ mm and then averaged at the scene and dataset levels. Lower values indicate more accurate focal-dependent geometry, with an ideal value of zero. The complete formulation is provided in Appendix B.

Following GP, we additionally report CorrCoef for focal-trajectory consistency. FOV-aligned DINOv2 evaluates cross-focal scene preservation, while CLIP measures text-image alignment. For all datasets and compared methods, CorrCoef is computed using the analytic focal trajectory derived from each method's own generated 24 mm anchor. This unified protocol evaluates relative focal-response consistency while avoiding interference from method-specific anchor-content differences. CorrCoef is averaged over valid scenes only.

Implementation details. Recovery-LF fine-tunes the camera encoder and the native camera-conditioned attention merge projections while keeping the remaining UNet, VAE, and text encoder frozen. Physics/G is subsequently optimized for 500 steps on 96 development scenes using AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. During Physics/G training, the recovered focal generator and VAE remain frozen. High-Frequency Appearance Preservation is used only during Physics/G training and introduces no additional inference-time parameters or processing. GP-based models use 25 DDIM sampling steps with classifier-free guidance scale 8.

B Evaluation Metric Details

B.1 Low-Frequency Geometry Gap

We introduce **LF Gap (Low-Frequency Geometry Gap)** to measure whether the focal-dependent geometric response of a generated focal sequence agrees with the corresponding analytic camera

transformation. The metric focuses on large-scale composition, image scale, and structural displacement while suppressing high-frequency appearance variations such as texture, sharpness, and color differences.

Analytic focal reference. Let the input focal sequence be

$$\mathcal{F} = \{f_0, f_1, \ldots, f_K\},$$

where f_0 denotes the anchor focal length. For each evaluated method, its own generated anchor image I_{f_0} is used to construct an analytic reference for each target focal length f_k.

Given sensor dimensions w and h, the horizontal and vertical fields of view are defined as

$$\theta_x(f) = 2\arctan\left(\frac{w/2}{f}\right), \qquad \theta_y(f) = 2\arctan\left(\frac{h/2}{f}\right).$$

The center-crop ratio for target focal length f_k relative to the anchor is

$$r_k = \min_{d\in\{x,y\}} \frac{\theta_d(f_k)}{\theta_d(f_0)}.$$

The corresponding analytic reference is obtained by center-cropping the anchor image according to r_k and resizing it back to the original resolution:

$$R_{f_k} = \mathrm{Resize}\left(\mathrm{CenterCrop}\left(I_{f_0}, r_k\right)\right).$$

In the current implementation, a 36×24 mm sensor is used and $f_0 = 24$ mm. Lanczos interpolation is used to resize the cropped reference.

Low-frequency intensity gap. Both R_{f_k} and the generated target image I_{f_k} are converted to grayscale and normalized to $[0, 1]$. Gaussian low-pass filtering is then applied to suppress high-frequency appearance information, followed by spatial downsampling. In our implementation, the Gaussian standard deviation is $\sigma = 8$ and the low-frequency representations are resized to 48×32.

Let L_R and L_I denote the resulting low-frequency representations of the analytic reference and generated target image, respectively. For two flattened representations A and B, we compute the zero-mean normalized correlation

$$\rho(A, B) = \frac{(A - \bar{A})^\top (B - \bar{B})}{\|A - \bar{A}\|_2 \|B - \bar{B}\|_2}.$$

The low-frequency intensity gap is defined as

$$G_{\mathrm{int}} = 1 - \rho(L_R, L_I).$$

Low-frequency gradient gap. To additionally evaluate large-scale structural boundaries, we compute the Sobel gradient magnitude

$$M(A) = \sqrt{S_x(A)^2 + S_y(A)^2},$$

where S_x and S_y denote the horizontal and vertical Sobel operators. The corresponding gradient gap is

$$G_{\mathrm{grad}} = 1 - \rho\left(M(L_R), M(L_I)\right).$$

LF Gap. The LF Gap for target focal length f_k is defined as the average of the intensity and gradient gaps:

$$G_{\mathrm{LF}}(f_k) = \frac{G_{\mathrm{int}} + G_{\mathrm{grad}}}{2},$$

or equivalently,

$$G_{\mathrm{LF}}(f_k) = 1 - \frac{\rho(L_R, L_I) + \rho\left(M(L_R), M(L_I)\right)}{2}.$$

A lower LF Gap indicates that the generated target view follows the expected focal-dependent low-frequency geometry more closely. An ideal value of zero indicates perfect agreement under this metric.

Scene- and dataset-level aggregation. For each scene, LF Gap is averaged over all non-anchor focal views:

$$G_{\text{scene}} = \frac{1}{K} \sum_{k=1}^{K} G_{\text{LF}}(f_k).$$

The final dataset-level score is

$$G_{\text{dataset}} = \frac{1}{N} \sum_{s=1}^{N} G_{\text{scene}}^{(s)}.$$

The anchor view itself is excluded from the average because it is used to construct the analytic references for the remaining focal views. For the main Formal80 and COCO-val500 experiments, the input sequence is $\{24, 35, 45, 55, 70\}$ mm and LF Gap is therefore averaged over $\{35, 45, 55, 70\}$ mm. For focal-range robustness experiments, the same metric is applied to the actual non-anchor focal views of each evaluated trajectory rather than to a fixed set of focal lengths.

C Extended Component Analysis

We further analyze the contribution of each component in FocalBridge. Starting from the GP baseline, applying Physics/G alone reduces LF Gap from 0.1944 to 0.1648 and increases CorrCoef from 0.8757 to 0.9030. This result shows that converting the focal-dependent prior into an explicit generation-time geometry reference provides an independent improvement over the original GP focal-control pathway.

Recovery-LF independently reduces LF Gap from 0.1944 to 0.1474 and improves DINOv2 from 0.6209 to 0.6308. This indicates that adapting the focal-control pathway to self-generated diffusion states improves the stability of focal-dependent structure under free-running generation.

Combining Recovery-LF with Physics/G further reduces LF Gap to 0.1231 and increases CorrCoef to 0.9066. The result demonstrates that self-generated-state adaptation and explicit physical geometry correction address complementary sources of focal error.

Finally, adding High-Frequency Appearance Preservation yields the complete FocalBridge model, reaching 0.1222 LF Gap, 0.9108 CorrCoef, and 0.6349 DINOv2, while keeping CLIP nearly unchanged. The relatively small but consistent improvement indicates that High-Frequency Appearance Preservation mainly constrains unnecessary local appearance changes during geometry correction rather than acting as an additional focal-control mechanism.

D Focal-Range Robustness

We further evaluate FocalBridge under focal trajectories with different ranges. Three five-view focal trajectories are considered: $\{24, 28, 32, 36, 40\}$ mm, $\{24, 32, 40, 48, 56\}$ mm, and $\{24, 35, 47, 58, 70\}$ mm, corresponding to focal spans of 16 mm, 32 mm, and 46 mm, respectively.

For each trajectory, 24 mm serves as the anchor view. LF Gap is computed over the four non-anchor focal views and then averaged at the scene level. Thus, the evaluated target focal sets are $\{28, 32, 36, 40\}$ mm, $\{32, 40, 48, 56\}$ mm, and $\{35, 47, 58, 70\}$ mm for the short, medium, and wide trajectories, respectively.

For all methods and datasets, LF Gap and CorrCoef use the same own-anchor reference protocol. Specifically, each method's own generated 24 mm image is used to construct the analytic FoV references for the remaining focal views. This unified protocol avoids using a dataset-specific or method-specific external reference and enables consistent comparison across Formal80 and COCO-val500. Complete results are reported in Tables 3 and 4.

LF-Gap analysis. On Formal80, FocalBridge is statistically comparable to GP at the 16 mm span, with an LF-Gap difference of $+0.00152$ and a paired 95% confidence interval of $[-0.01203, 0.01597]$. At the 32 mm and 46 mm spans, FocalBridge reduces LF Gap by 0.01855 and 0.06922, respectively, with paired 95% confidence intervals of $[-0.03397, -0.00302]$ and

Table 3: Focal-range robustness on Formal80. Values are reported as mean with scene-level bootstrap 95% confidence intervals. CorrCoef additionally reports the number of valid scenes. Best results within each focal span are shown in bold.

Span	Method	LF Gap↓	CorrCoef↑	DINOv2↑	CLIP↑
Short 16 mm	GP	**0.0630** [0.0545, 0.0721]	0.7491 [0.6677, 0.8199], 79/80	**0.7384** [0.7311, 0.7457]	0.2899 [0.2818, 0.2980]
	CameraCtrl-FT-Focal	0.0841 [0.0684, 0.1021]	0.7579 [0.6857, 0.8220], 78/80	0.7227 [0.7141, 0.7312]	0.2937 [0.2866, 0.3007]
	FocalBridge	0.0645 [0.0526, 0.0778]	**0.8276** [0.7472, 0.8932], 77/80	0.7173 [0.7072, 0.7271]	**0.2938** [0.2864, 0.3011]
Medium 32 mm	GP	0.1268 [0.1111, 0.1440]	0.8524 [0.7818, 0.9093], 79/80	0.6587 [0.6474, 0.6698]	0.2876 [0.2803, 0.2948]
	CameraCtrl-FT-Focal	0.1362 [0.1190, 0.1547]	**0.9384** [0.8986, 0.9666], 77/80	0.6571 [0.6474, 0.6664]	**0.2925** [0.2855, 0.2996]
	FocalBridge	**0.1082** [0.0919, 0.1260]	0.9297 [0.9054, 0.9501], 79/80	**0.6605** [0.6512, 0.6700]	0.2860 [0.2793, 0.2926]
Wide 46 mm	GP	0.2032 [0.1809, 0.2273]	0.9174 [0.8488, 0.9680], 79/80	0.6145 [0.6034, 0.6257]	0.2872 [0.2806, 0.2937]
	CameraCtrl-FT-Focal	0.1996 [0.1795, 0.2210]	**0.9227** [0.8587, 0.9712], 79/80	0.6119 [0.6009, 0.6228]	**0.2928** [0.2857, 0.2999]
	FocalBridge	**0.1340** [0.1174, 0.1516]	0.8959 [0.8186, 0.9573], 77/80	**0.6289** [0.6176, 0.6403]	0.2882 [0.2813, 0.2953]

Table 4: Focal-range robustness on COCO-val500. Values are reported as mean with scene-level bootstrap 95% confidence intervals. CorrCoef additionally reports the number of valid scenes. Best results within each focal span are shown in bold.

Span	Method	LF Gap↓	CorrCoef↑	DINOv2↑	CLIP↑
Short 16 mm	GP	**0.0676** [0.0629, 0.0726]	0.6750 [0.6417, 0.7069], 498/500	**0.7440** [0.7412, 0.7469]	0.2806 [0.2778, 0.2833]
	CameraCtrl-FT-Focal	0.1167 [0.1087, 0.1251]	0.8026 [0.7760, 0.8274], 495/500	0.7129 [0.7088, 0.7169]	0.2791 [0.2762, 0.2819]
	FocalBridge	0.0816 [0.0761, 0.0873]	**0.8468** [0.8216, 0.8696], 494/500	0.7153 [0.7115, 0.7192]	**0.2825** [0.2798, 0.2852]
Medium 32 mm	GP	0.1373 [0.1311, 0.1437]	0.9066 [0.8888, 0.9221], 497/500	**0.6650** [0.6613, 0.6687]	**0.2806** [0.2779, 0.2833]
	CameraCtrl-FT-Focal	0.1762 [0.1675, 0.1852]	0.8968 [0.8710, 0.9203], 497/500	0.6499 [0.6460, 0.6538]	0.2783 [0.2756, 0.2810]
	FocalBridge	**0.1155** [0.1089, 0.1224]	**0.9231** [0.9054, 0.9386], 493/500	0.6619 [0.6579, 0.6657]	0.2790 [0.2763, 0.2818]
Wide 46 mm	GP	0.2202 [0.2115, 0.2291]	**0.8918** [0.8627, 0.9181], 495/500	0.6142 [0.6099, 0.6184]	0.2788 [0.2762, 0.2814]
	CameraCtrl-FT-Focal	0.2625 [0.2525, 0.2729]	0.8558 [0.8206, 0.8883], 495/500	0.6019 [0.5980, 0.6058]	0.2791 [0.2763, 0.2818]
	FocalBridge	**0.1575** [0.1505, 0.1646]	0.8814 [0.8524, 0.9078], 492/500	**0.6270** [0.6230, 0.6310]	**0.2793** [0.2767, 0.2820]

$[-0.08667, -0.05327]$. The improvement therefore becomes substantially larger as the focal range increases.

A similar span-dependent trend is observed on COCO-val500. At the 16 mm span, FocalBridge has a higher LF Gap than GP by $+0.01405$, with a paired 95% confidence interval of $[0.00705, 0.02099]$. At the 32 mm and 46 mm spans, however, FocalBridge reduces LF Gap by 0.02181 and 0.06267, respectively, with paired 95% confidence intervals of $[-0.02865, -0.01485]$ and $[-0.07017, -0.05513]$. These results consistently show that the benefit of explicit physical geometry correction becomes more pronounced for medium-to-large focal changes.

CorrCoef analysis. CorrCoef exhibits a different pattern from LF Gap. Under the unified own-anchor protocol, FocalBridge does not achieve the highest CorrCoef for every focal range, indicating that trajectory correlation and absolute focal-geometry accuracy capture complementary properties. Existing camera-conditioned methods can maintain a highly correlated focal response while still deviating from the expected physical magnitude. LF Gap directly measures this absolute low-frequency geometry deviation and therefore provides a complementary measure of focal-control accuracy.

Scene preservation. At the widest 46 mm range, FocalBridge achieves higher DINOv2 than GP on both Formal80 (0.6289 vs. 0.6145) and COCO-val500 (0.6270 vs. 0.6142). Thus, the reduction in focal-geometry error at large focal spans is not accompanied by degraded cross-focal scene preservation. Overall, these results indicate that FocalBridge is particularly effective for medium-to-large focal changes, where focal-dependent geometric deviations become more pronounced.

E Efficiency and Complexity

Let N denote the number of standard denoising steps and F_t the number of non-anchor focal views. The dominant inference complexity is

$$\mathcal{C}_{\mathrm{FB}} = \mathcal{O}(N\mathcal{C}_U + F_t\mathcal{C}_G + F_t\mathcal{C}_E)\,, \tag{15}$$

where $\mathcal{C}_U$ denotes one joint multi-focal denoising step, $\mathcal{C}_G$ denotes one geometry-controlled VAE decoding operation, and $\mathcal{C}_E$ denotes the physical-reference encoding cost.

FocalBridge introduces no additional diffusion evaluations compared with the underlying GP sampling process. For a five-view focal sequence, the complete FocalBridge pipeline requires approximately 8.36 s, compared with 7.65 s for GP under the same evaluation setting.

F Additional Qualitative Results

We provide additional qualitative focal sequences to complement the examples shown in the main paper. All compared methods use identical prompts, focal settings, random seeds, and output resolutions within each scene.

The additional examples cover different scene types and focal ranges. For each example, generated focal sequences can be presented together with their corresponding low-frequency focal-error maps. These examples provide further qualitative evidence of focal-dependent geometry behavior beyond the representative Formal80 cases shown in the main paper.

G Additional Theoretical Details

For completeness, we provide additional details for the bounded geometry correction used in Physics/G. Let the injected decoder residual be

$$\Delta H_i = b_i \tanh\left(\frac{\Delta H_i^{\mathrm{raw}}}{b_i}\right), \qquad b_i = \lambda_b \max\left(\mathrm{RMS}(H_i^{\mathrm{mid}}), \varepsilon\right).$$

Since $|\tanh(\cdot)| \le 1$, every element of the injected correction is bounded, and therefore

$$\|\Delta H_i\|_\infty \le b_i.$$

Let D_{up} denote the remaining frozen VAE decoder after the mid-block. If D_{up} is L_D-Lipschitz, then

$$\|I_i^{\mathrm{phy}} - I_i^{\mathrm{base}}\|_2 \le L_D \|\Delta H_i\|_2,$$

where

$$I_i^{\mathrm{base}} = D_{\mathrm{up}}(H_i^{\mathrm{mid}}),$$

and

$$I_i^{\mathrm{phy}} = D_{\mathrm{up}}\left(H_i^{\mathrm{mid}} + \Delta H_i\right).$$

Using the element-wise bound above gives

$$\|\Delta H_i\|_2 \le \sqrt{CHW}\, b_i,$$

and consequently

$$\|I_i^{\mathrm{phy}} - I_i^{\mathrm{base}}\|_2 \le L_D \sqrt{CHW}\, b_i.$$

The bounded residual therefore limits the magnitude of decoder-space geometry correction relative to the current VAE representation. High-Frequency Appearance Preservation further constrains unnecessary local appearance changes during training while introducing no additional inference-time operation.