

マルチモーダル潜在空間を通じた 少数視点合成のための予測的意味正規化

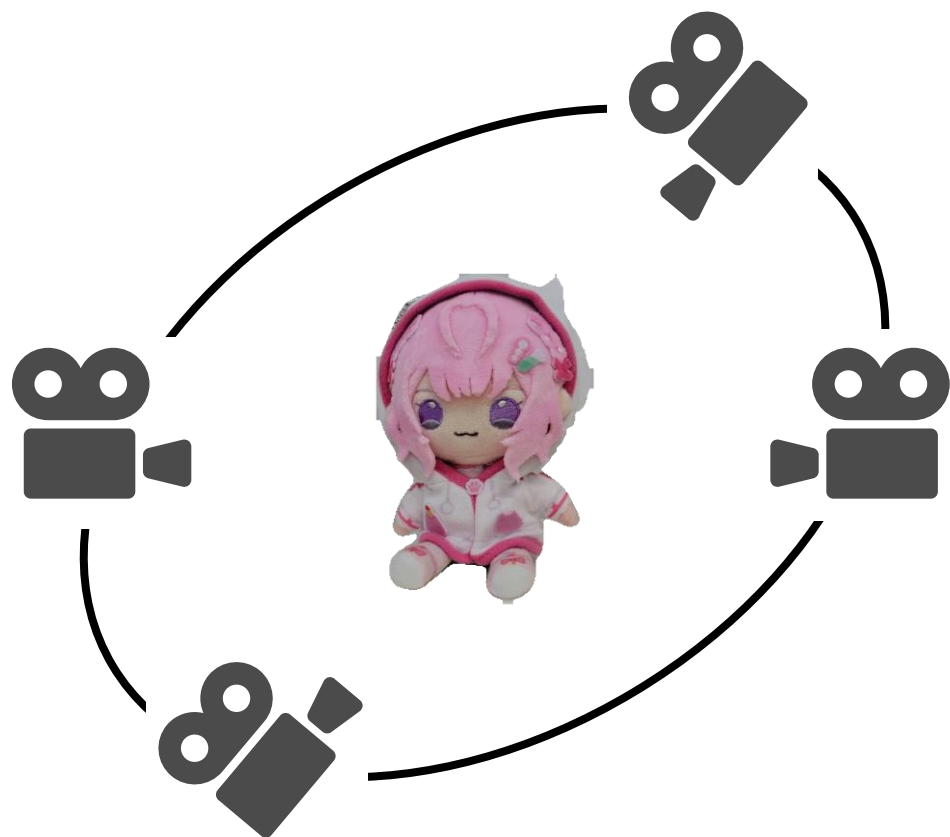
電気通信大学 柳井研究室

松浦史明

はじめに：自由視点映像再構成

3次元再構成（3D Reconstruction）

- ・ 物体を様々な角度から撮影した画像を入力し、レンダラのパラメータを学習
- ・ レンダラを用いて、任意の視点からの映像を出力



レンダラ



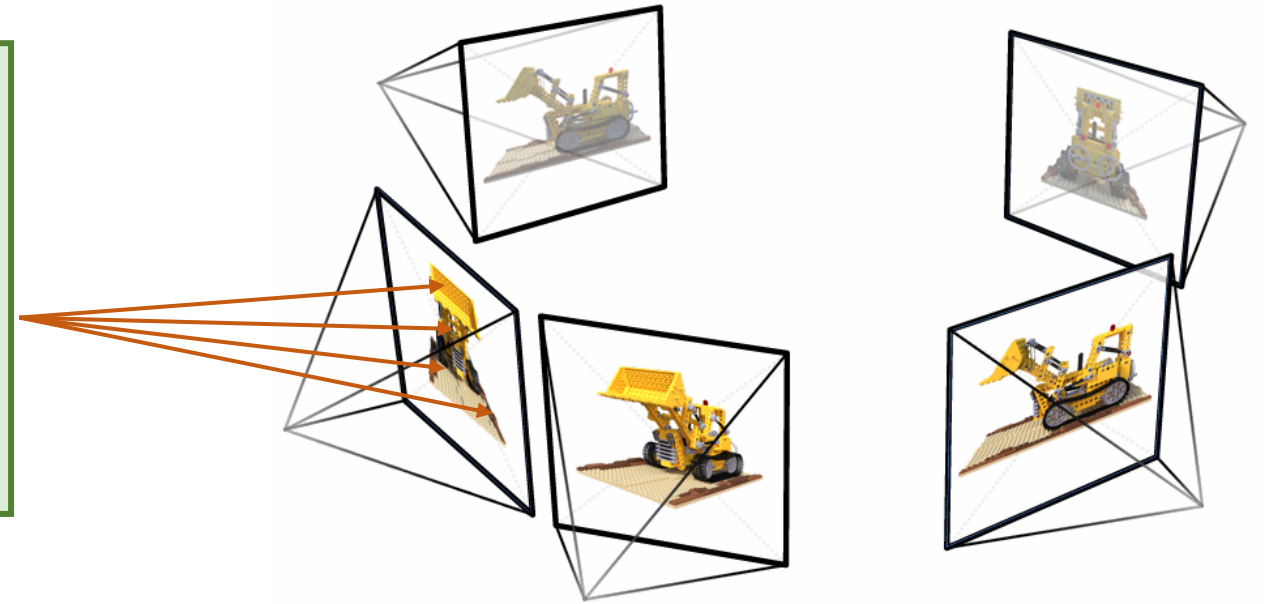
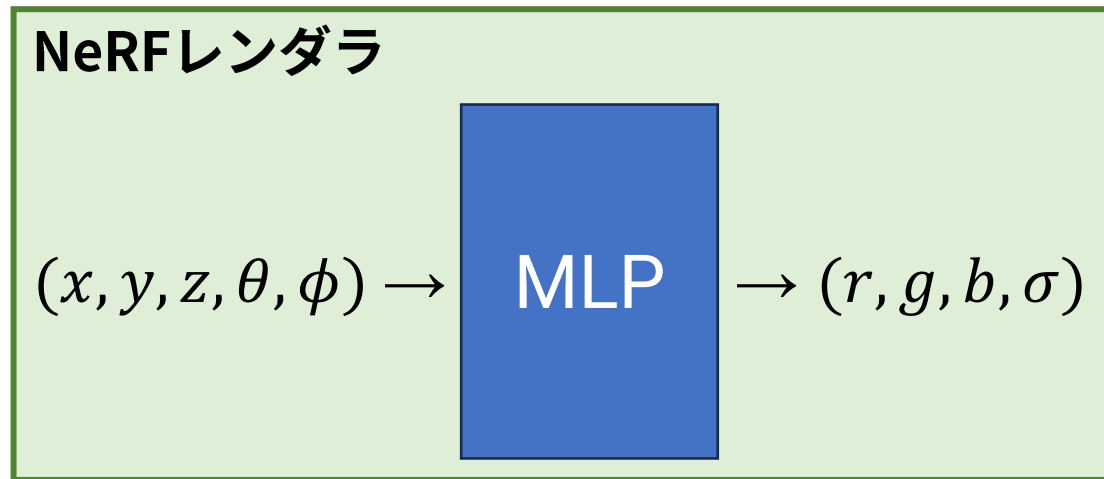
自由視点映像を連続再生した様子

はじめに：3次元表現

Neural Radiance Fields (NeRF) [4]

- ・ レンダラとして、ニューラルネット+ボリュームレンダリングを採用
- ・ 任意視点の2D画像を生成 (3Dを生成はしない)

NeRF

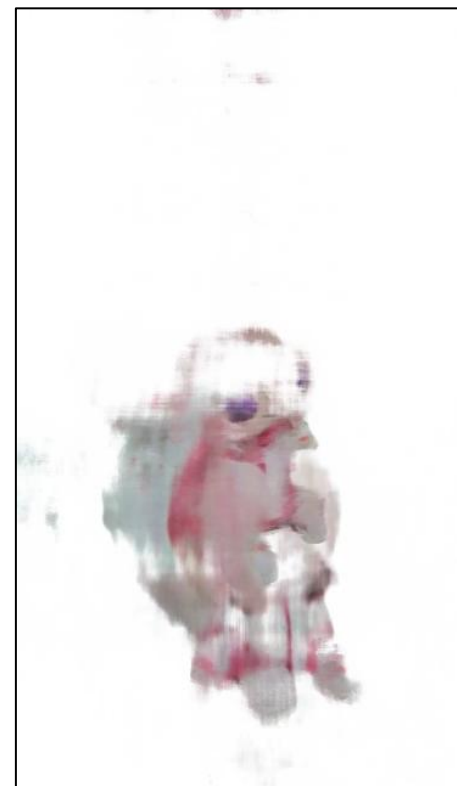
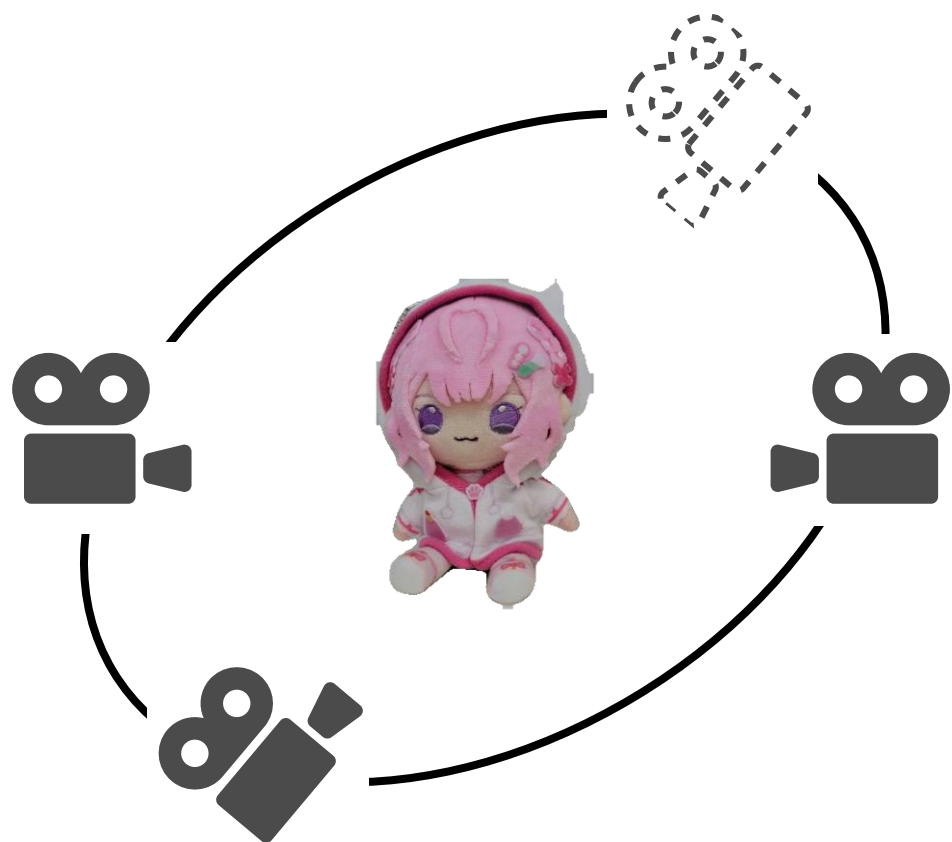


Images + accurate camera poses

はじめに：自由視点映像再構成

■ 少数ショット, 偏った視点などの厳しい入力条件

- ・ 未観測領域に起因する、再構成精度の大幅な低下

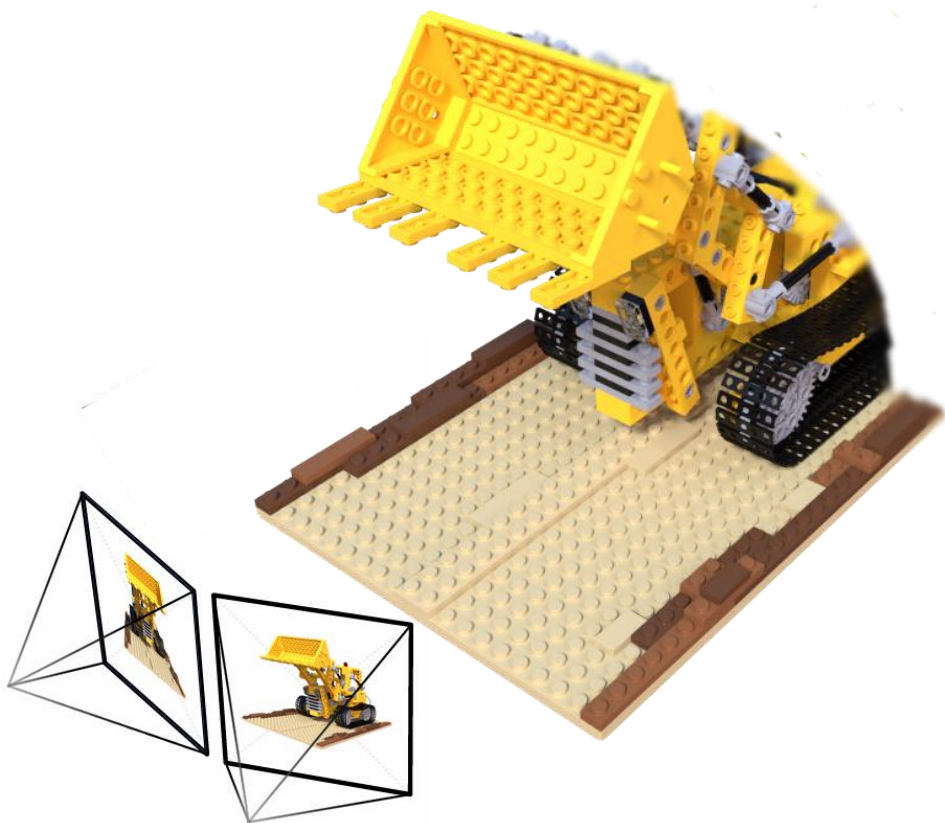


自由視点映像を連続再生した様子

はじめに：背景と目的

背景

- 未観測領域を含むデータからの再構成
⇒ アーティファクトの発生



目的

- 未観測領域を含む、完全な3次元再構成の実現

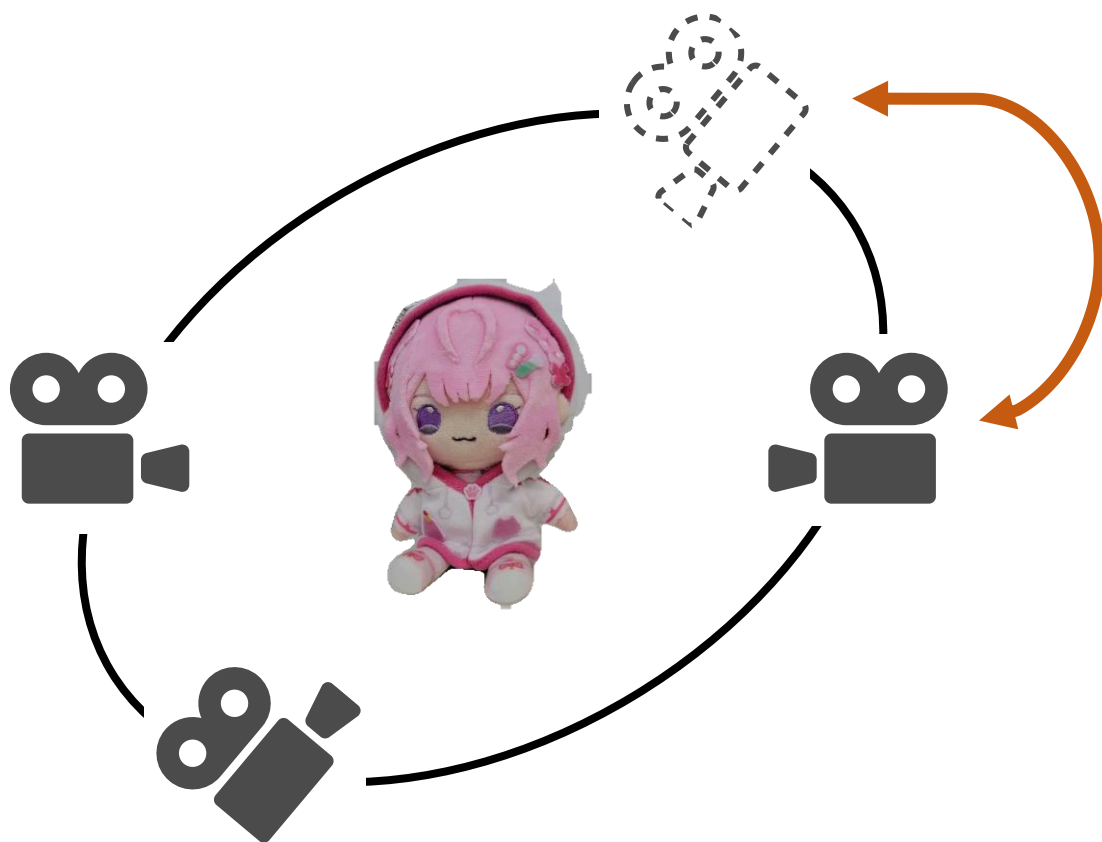


関連研究

関連研究：3次元再構成（NeRF）に係る、種々の正則化

レンダリング品質の改善のための、種々の正則化

- ・ 学習時の、入力表現や学習過程・損失関数の工夫
- ・ 既存手法では、正則化追加や拡散モデル依存が多い



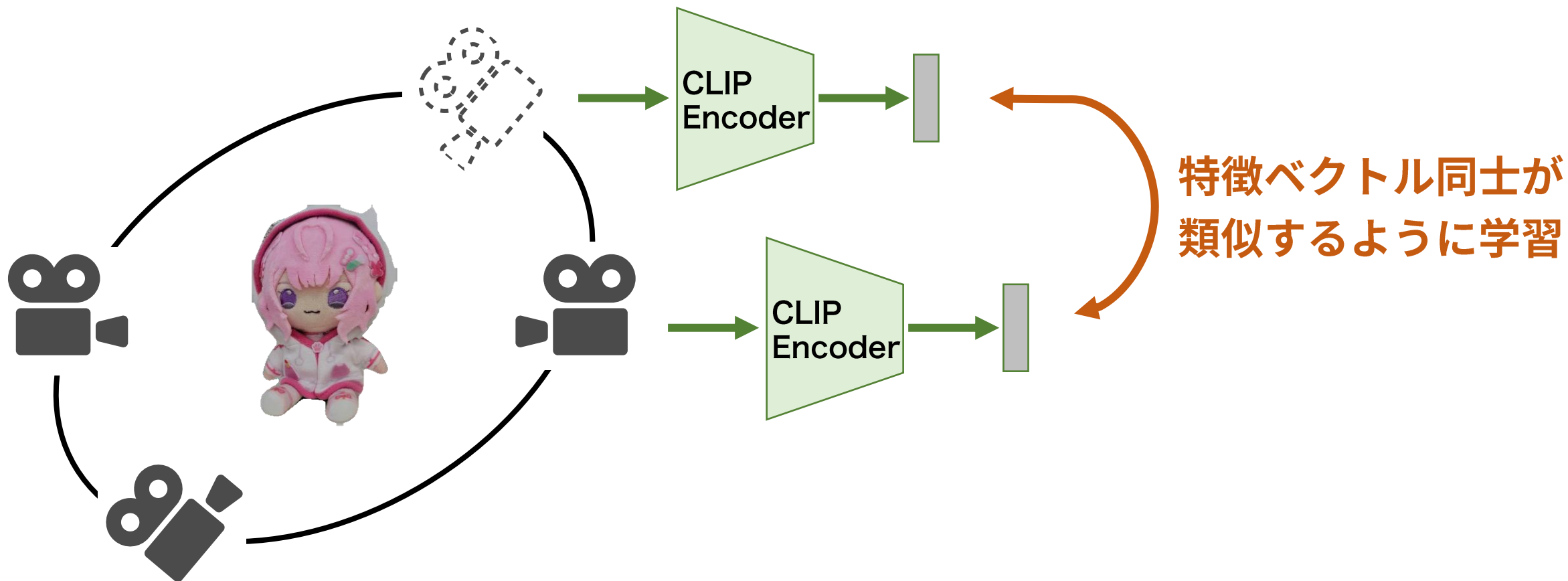
- ・ 深度平滑化
- ・ 色尤度の最大化
- ・ Positional Encoding の周波数制御
- ・ 相互情報量の最大化
- ・ 拡散モデルによる画像・潜在空間生成

など

関連研究：3次元再構成（NeRF）に係る、種々の正則化

DietNeRF [3]

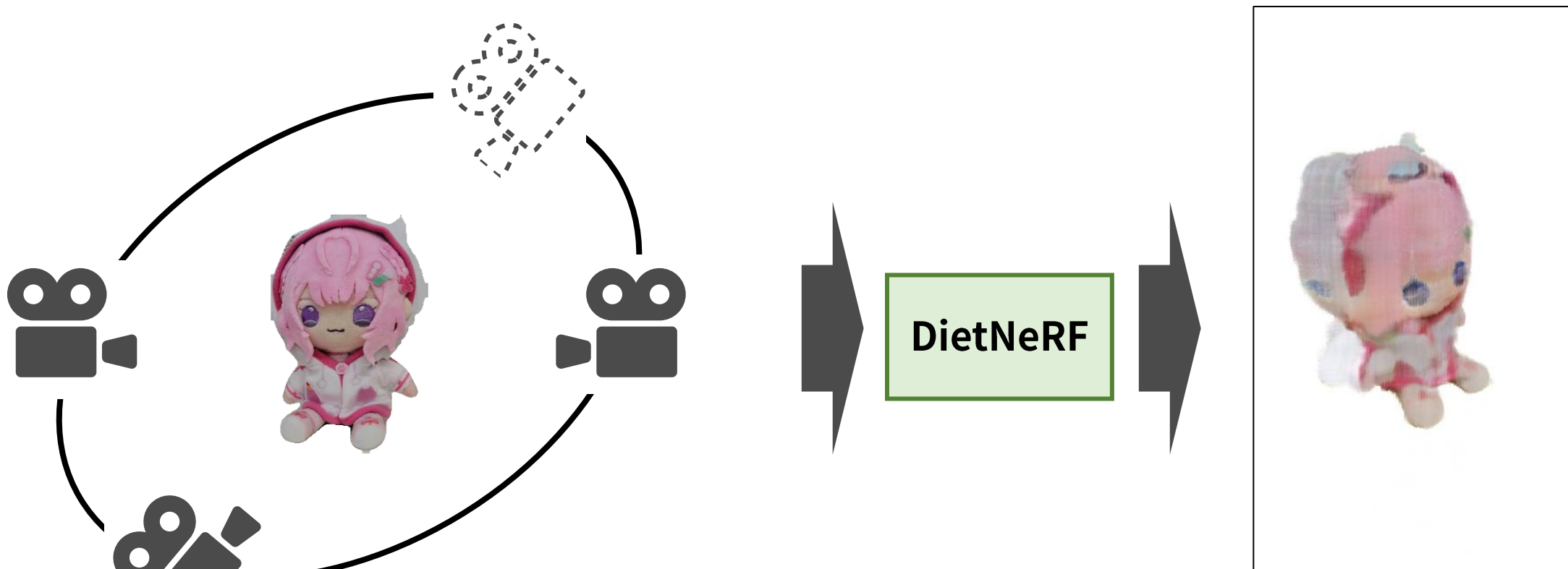
- ・ CLIP特徴類似度の最大化
- ・ 未観測領域であっても、同一物体であれば同一意味になるという考え方より



関連研究：3次元再構成（NeRF）に係る、種々の正則化

DietNeRF [3]

- ・ CLIP特徴類似度の最大化
- ・ 未観測領域であっても、同一物体であれば同一意味になるという考え方より



自由視点映像を連続再生した様子

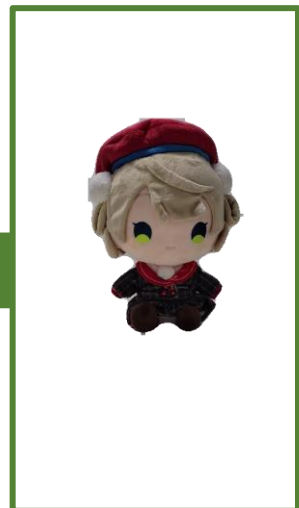
補足：2次元媒体を通した、自由視点映像生成の限界



Prompt: A video of a character plush toy wrapped around from the front to the back, like the attached image when viewed from the front

補足：大規模3D生成モデルの限界

Input Image



SAM3D (Meta)



Ground Truth

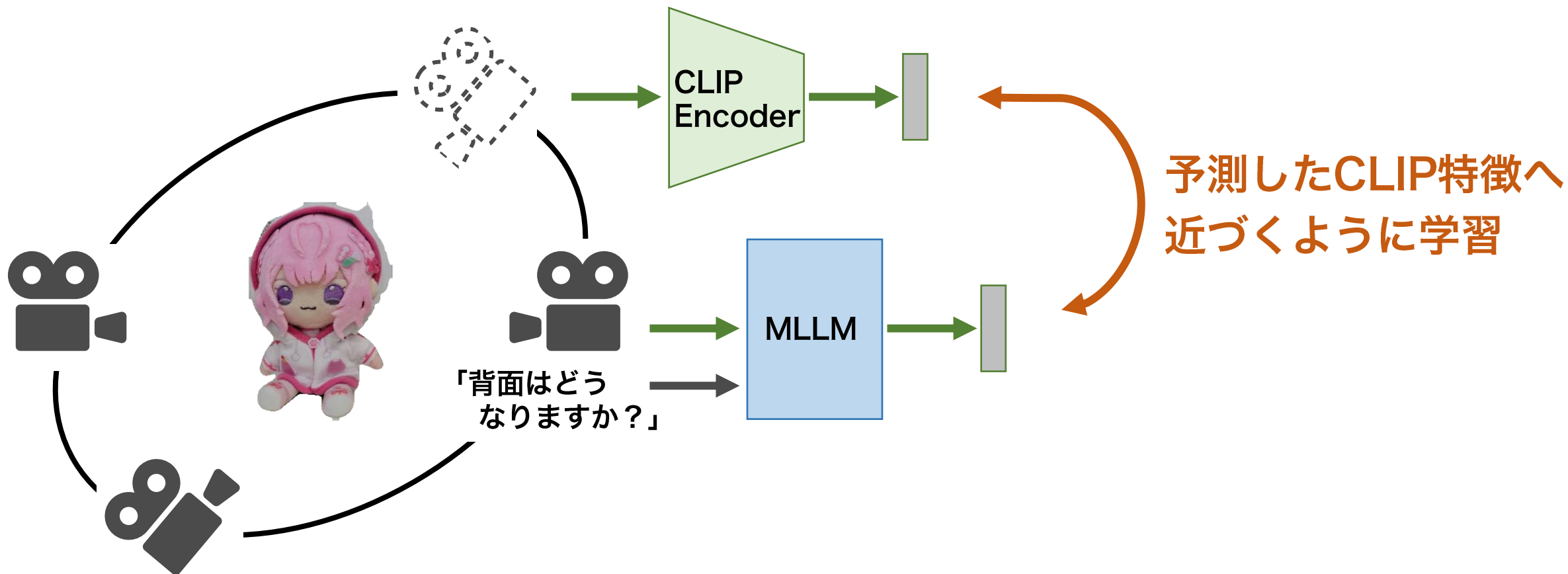


提案手法

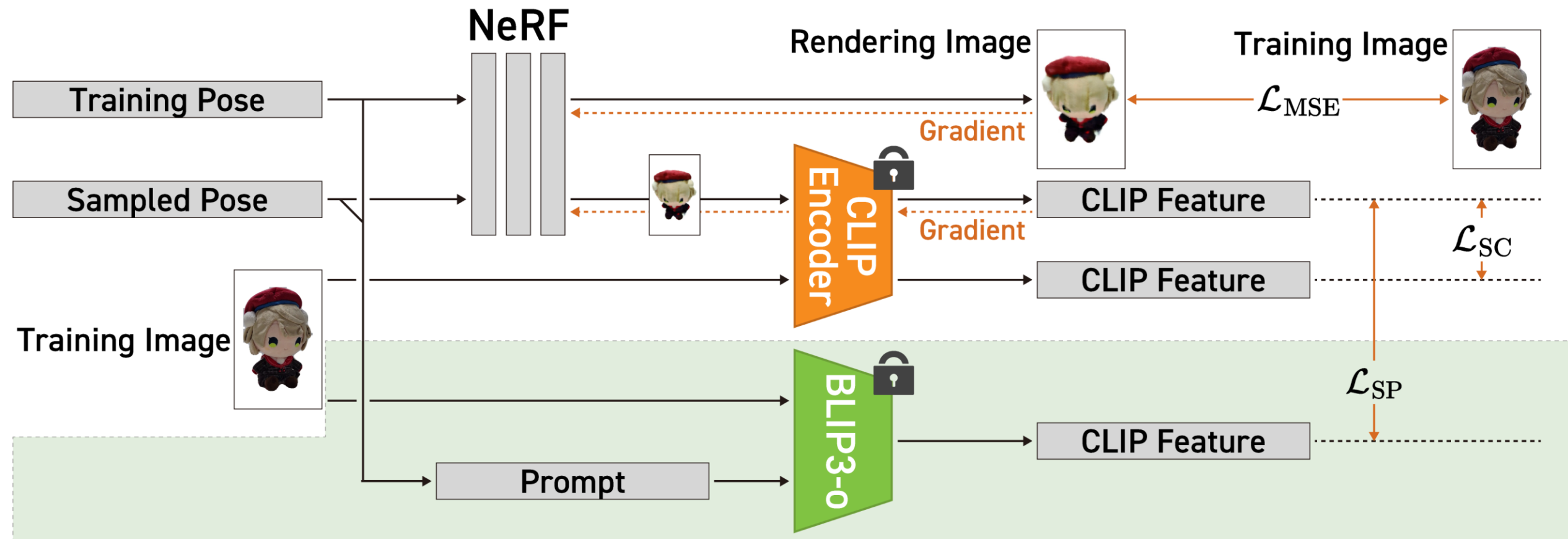
提案手法: 全体像

提案手法

- ・ 観測画像と相対回転をプロンプトに埋め込み、未観測領域の意味特徴を推定
- ・ NeRFレンダリング結果の特徴を、その予測特徴へ近づくよう学習を誘導



提案手法: モデル構造の概要

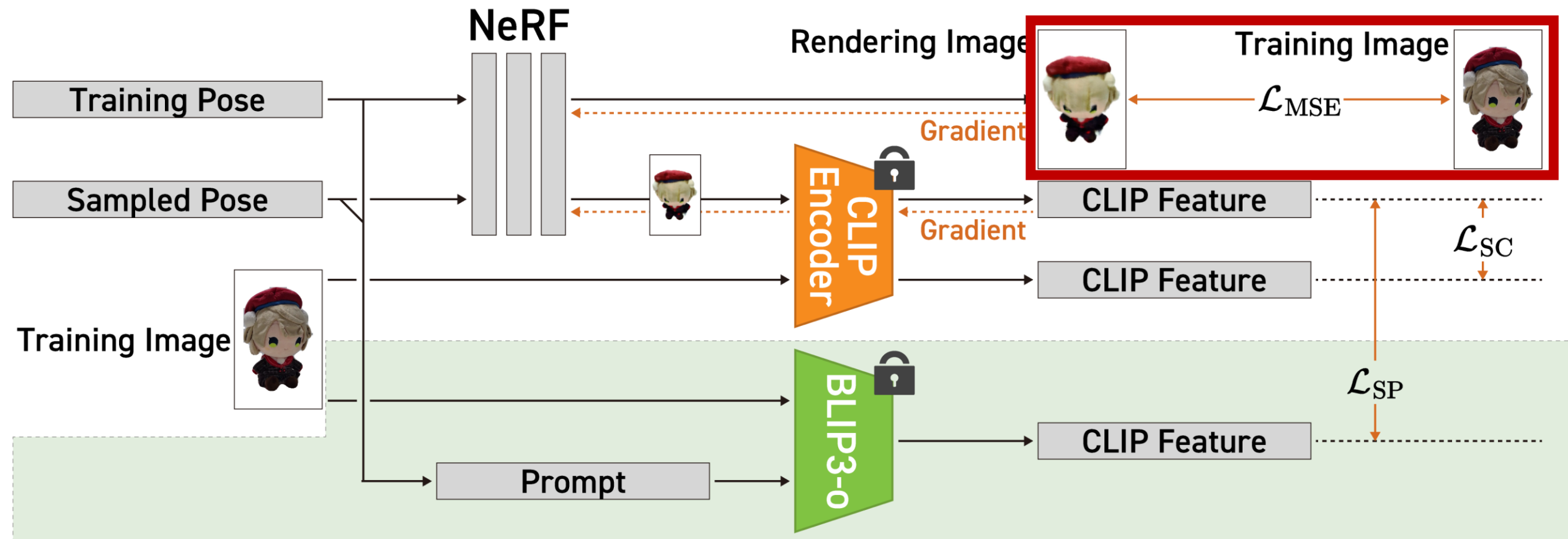


Added Component

3種の損失関数

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{sc}} \mathcal{L}_{\text{SC}, l_2} + \lambda_{\text{sp}} \mathcal{L}_{\text{SP}, l_2}$$

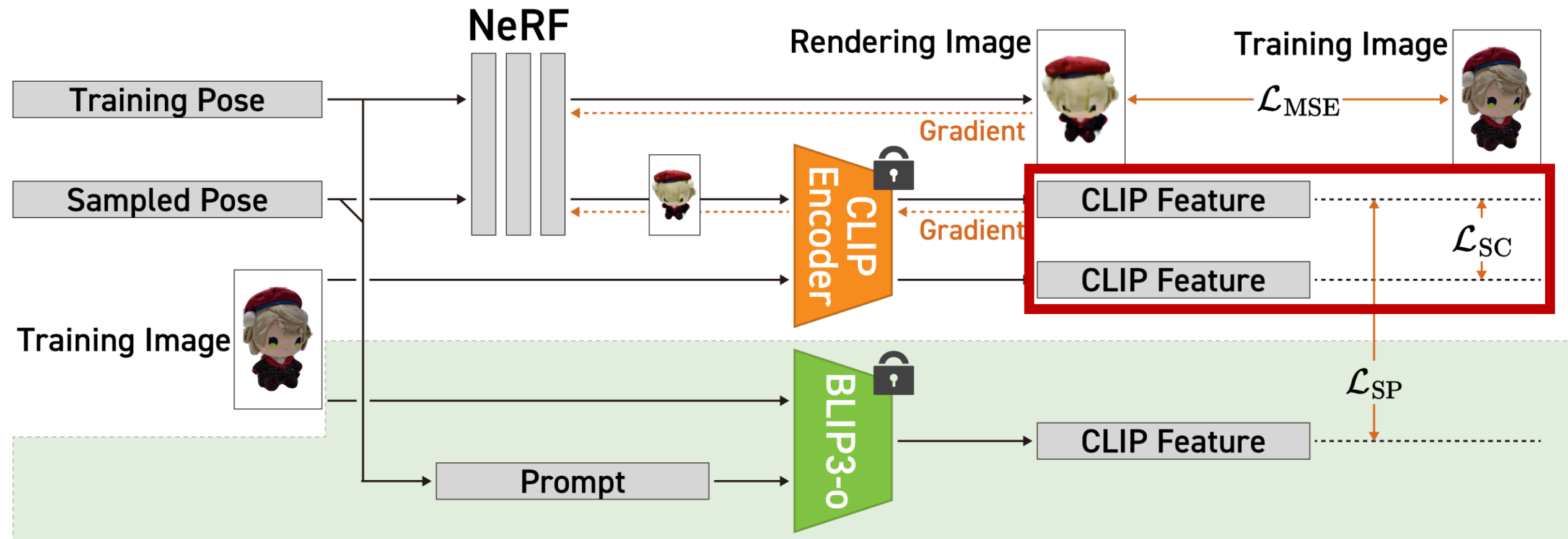
提案手法: 損失関数 (レンダリング損失) [4]



Added Component

$$\mathcal{L}_{MSE} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \left\| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\|_2^2$$

提案手法: 損失関数 (意味一貫性損失) [3]

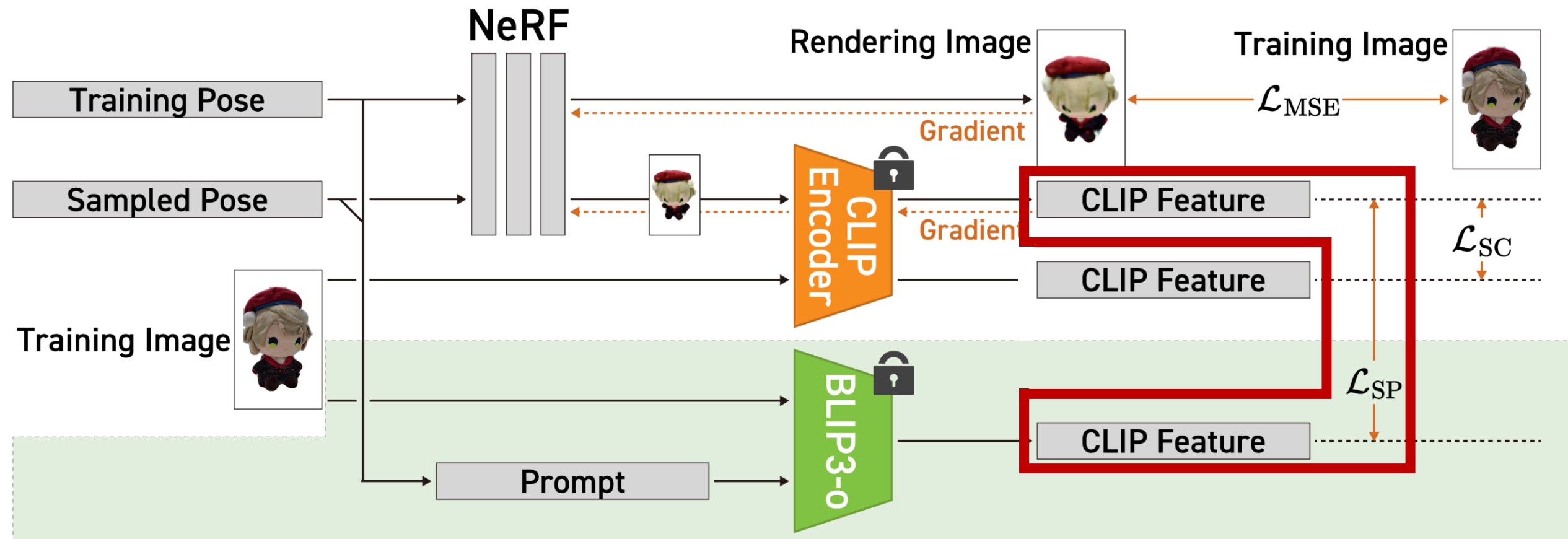


Added Component

$$\mathcal{L}_{SC, l_2} \left(I, \hat{I}_{\mathbf{p}_r} \right) = \lambda_{sc} \left\| \phi \left(I \right) - \phi \left(\hat{I}_{\mathbf{p}_r} \right) \right\|_2^2$$

提案手法: 損失関数 (意味予測損失)

【新規提案】



Added Component

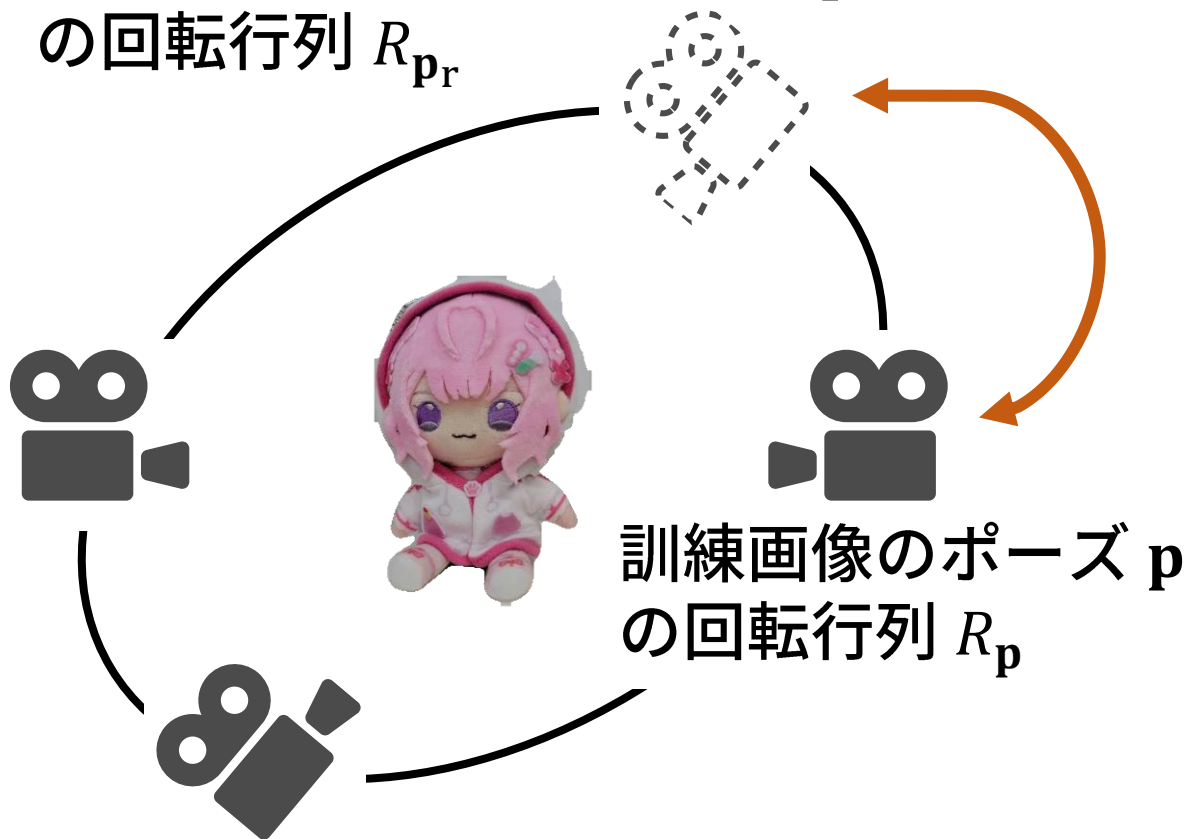
$$\mathcal{L}_{SP, \ell_2} \left(I, \hat{I}_{\mathbf{p}_r} \right) = \lambda_{sp} \left\| B(I, prompt) - \phi \left(\hat{I}_{\mathbf{p}_r} \right) \right\|_2^2$$

提案手法: プロンプト **【新規提案】**

日本語訳

オブジェクトを右に○度、上に○度、視点を○度回転させたときの、全体的な外観を想像してください...

レンダリング画像のポーズ p_r
の回転行列 R_{p_r}

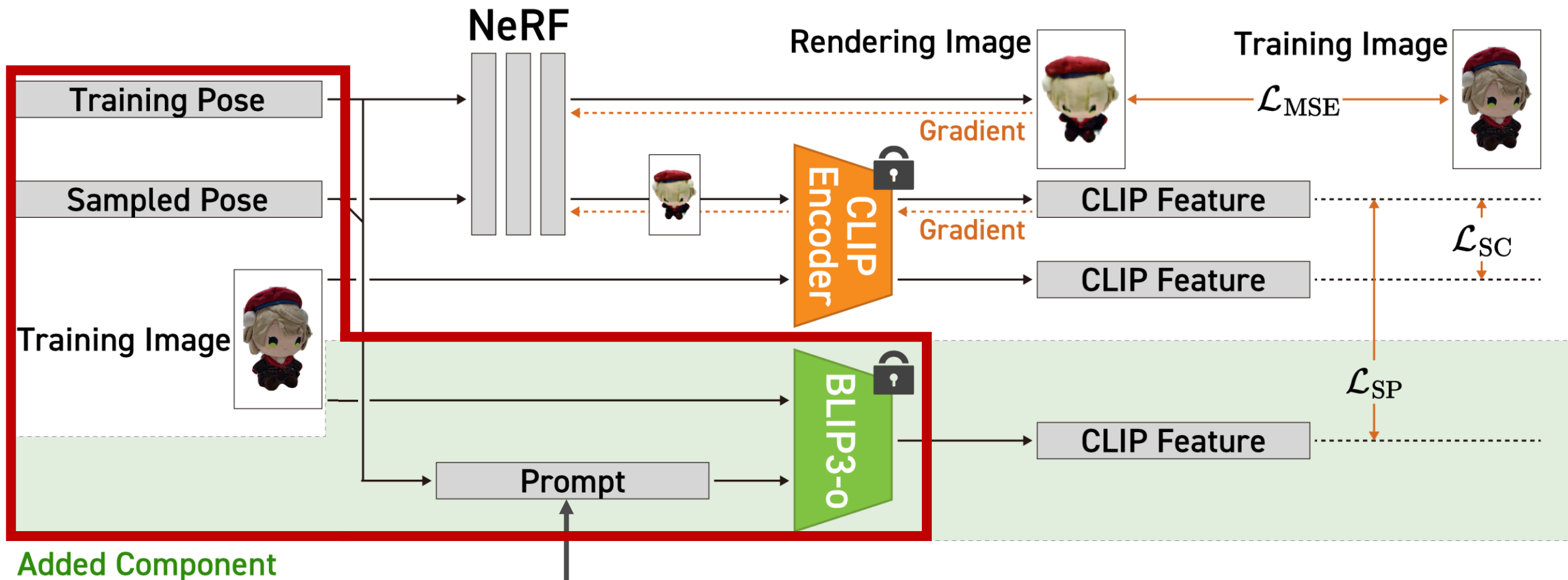


相対回転行列 $R_{rel} = R_p R_{p_r}^T$



オイラー角 (yaw, pitch, roll)

提案手法: プロンプト 【新規提案】



元プロンプト

Please generate image based on the following caption: Envision the central object rotated by {yaw:.0f} $^{\circ}$ to the right, {pitch:.0f} $^{\circ}$ upward, and {roll:.0f} $^{\circ}$ clockwise; describe its overall appearance –including shape, color, and texture—and specify any objects or features that would be absent, hidden, or not included from this viewpoint.

日本語訳

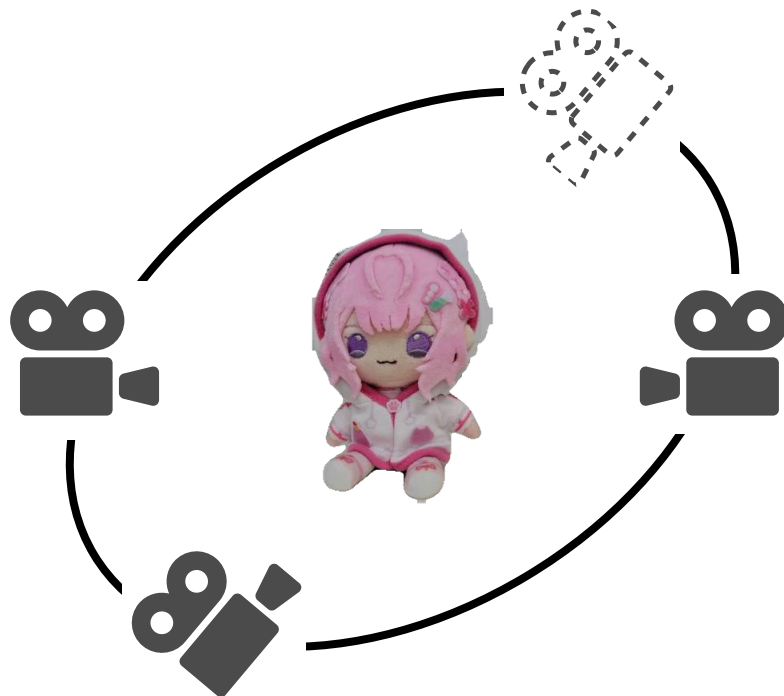
対象物体を右に○度、上に○度、視点を○度回転させたときの、全体的な外観を想像してください…

実験

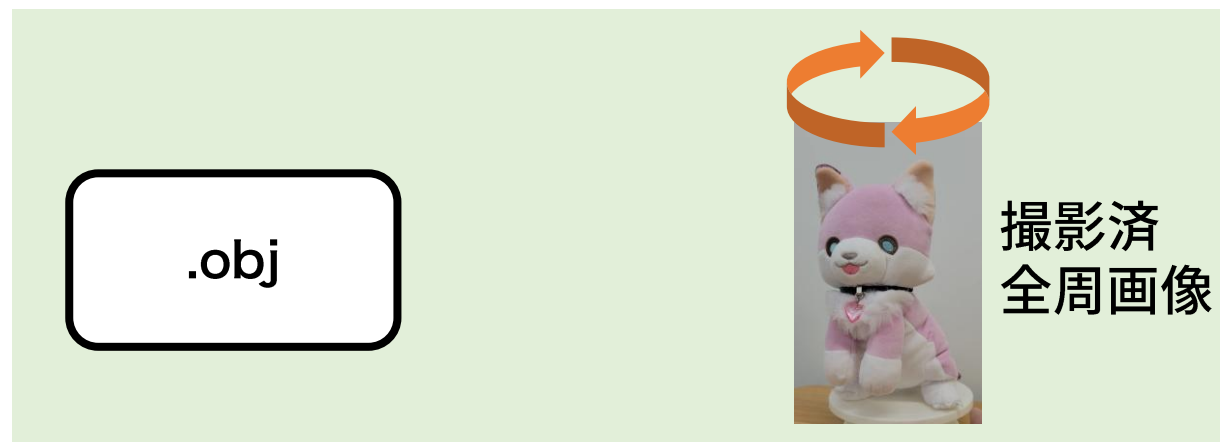
実験: データセット

データセット

- ・ 独自に作成した10種類 + GSOデータセット[2] より18種類の全周画像
- ・ 視点ごとランダムに、train / validation / test データに設定
 - ・ train / validation については背面画像を除き、背面の整合性を取りながら再構築可能かどうかを、testデータを用いて評価



実験: データセット



レンダリング



セグメンテーション



Train / Validation に用いる視点範囲

Test に用いる視点範囲

実験: 評価指標

PSNR

$$\text{PSNR} = 10 \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right)$$

- ・ ピクセル単位での一致度を測る

SSIM

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

- ・ ブロック単位での構造的類似度の平均

LPIPS

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0,hw}^l)\|_2^2$$

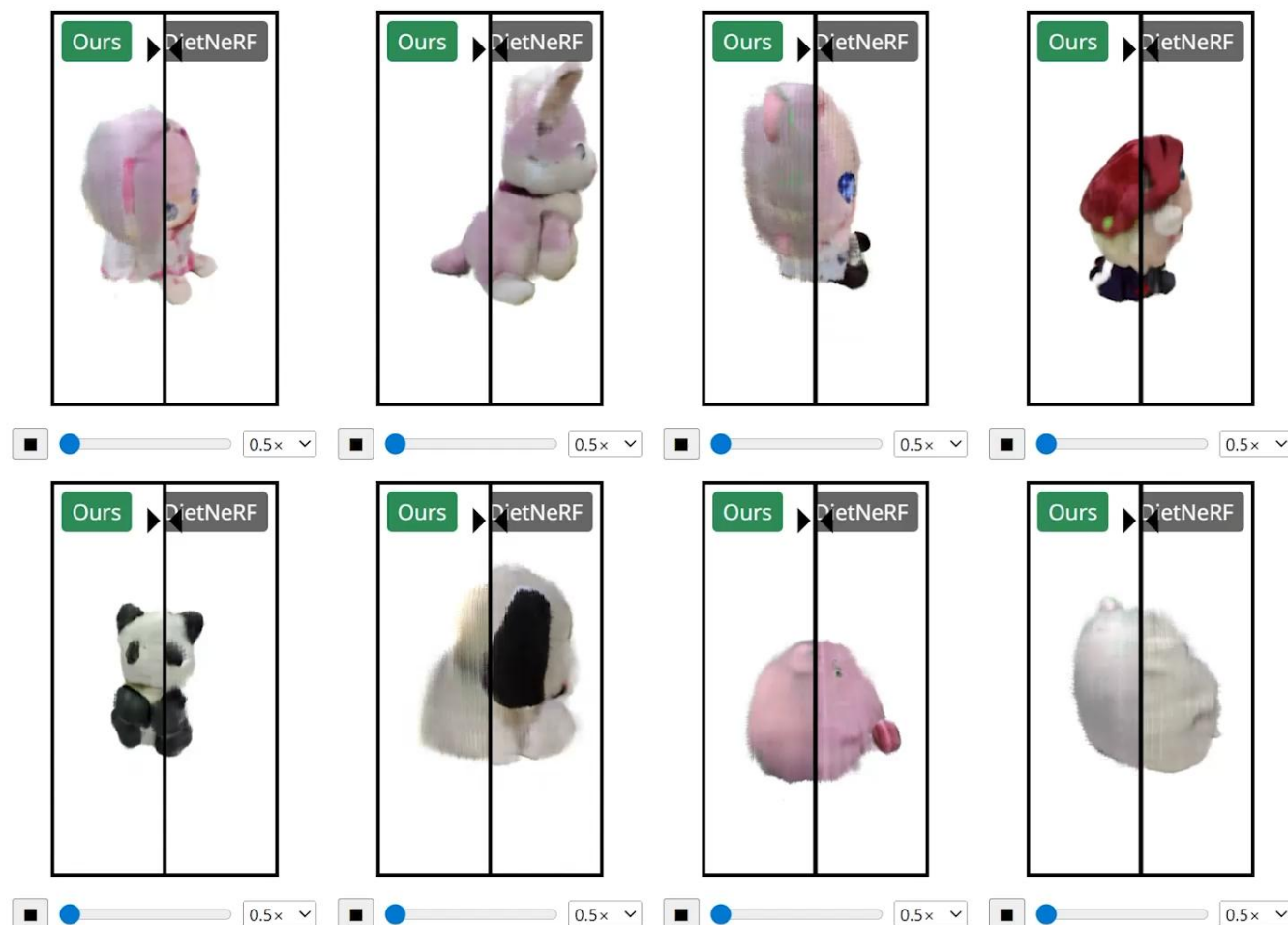
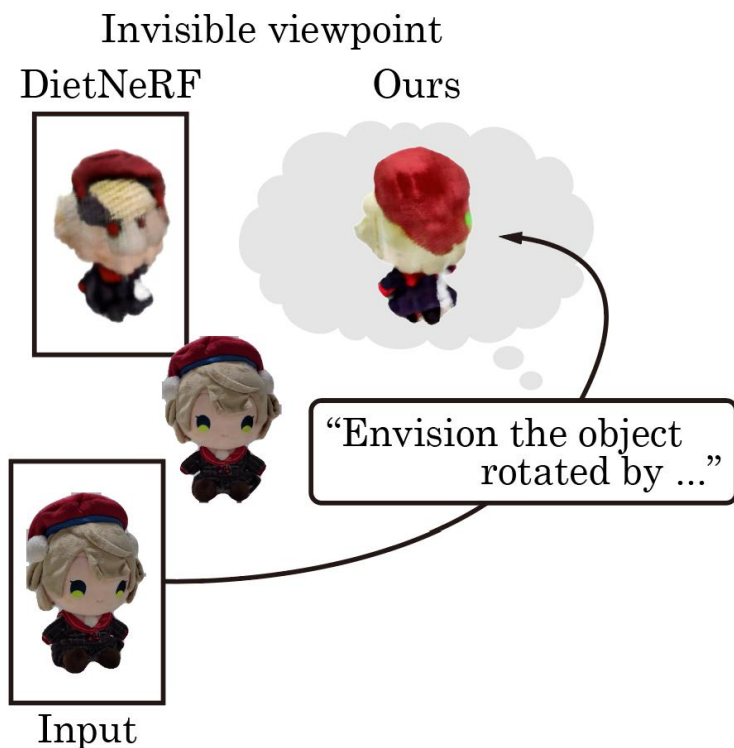
- ・ 学習済モデル (VGG) の中間層から得られる特徴量間の距離を計算
- ・ 人間の知覚的な評価と高い相関を持つ

実験: 実験設定

Method	#Train	Image Encoder ϕ	λ_{sc}	$Freq_{sc}$	λ_{sp}	$Freq_{sp}$
NeRF	8	-	-	-	-	-
DietNeRF	8	CLIP-ViT-B-32	0.1	10	-	-
DietNeRF (eva-clip)	8	EVA-CLIP-E-14-plus	0.1	10	-	-
Ours	8	EVA-CLIP-E-14-plus	0.1	10	0.05	10

- ・ #Train: 学習用画像から実際に学習に用いる画像の枚数
- ・ $Freq_{sc}, Freq_{sp}$: 学習イテレーションでの意味一貫性損失と意味予測損失の適用間隔
- ・ 学習イテレーション: 20,000回
- ・ 意味予測損失に関連する箇所以外はすべて、DietNeRF [3] に準ずる










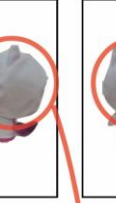

























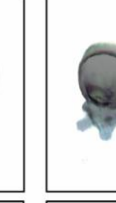

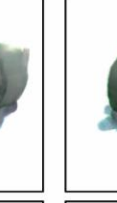
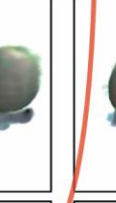
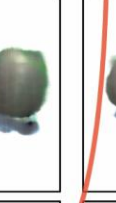
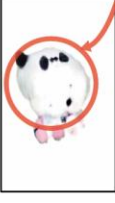









実験結果：定性評価



松浦史明, 中溝雄斗, 田邊光, 柳井啓司. BlipNeRF: MLLM による不可視視点の予測を用いた少数の物体正面画像からの3次元再構成.
画像の認識・理解シンポジウム (MIRU), 2025.

プロジェクトページ: [BlipNeRF: Semantically Predicted Few-Shot View Synthesis from Biased Views](#)

実験結果：定性評価

	Image pose range for training									
GT (Similar pose)										
DietNeRF										
OURS										
$\mathcal{L}_{SP} + \mathcal{L}_{MSE}$										
$\mathcal{L}_{SC} + \mathcal{L}_{SP} + \mathcal{L}_{MSE}$ (Changed in the prompt)										

実験結果：定量評価

評価指標の平均値

- ・ 従来手法であるDietNeRFと比較して、知覚的類似度が向上
- ・ 一方、意味の優先とピクセルレベルでの品質の両立は難しい

Method	PSNR↑	SSIM↑	LPIPS↓
NeRF	24.9	0.925	0.100
DietNeRF	25.6	0.924	0.094
DietNeRF (eva-clip)	23.5	0.922	0.098
Ours	24.2	0.927	0.092

再構成に要する時間

RTX A6000を2台用いた場合（1つのデータセットあたり）

- ・ NeRF: 約1時間30分
- ・ DietNeRF: 約2h50分
- ・ DietNeRF (eva-clip) : 約3時間
- ・ 提案手法: 約4時間40分

※NeRF単体で4GB程度、CLIP や EVA-CLIP で10GB超、BLIP3-o 単体で40GB超のVRAMを要する

実験結果：アブレーション分析

MLLMの理解力の確認

Envision the central object rotated by 0° to the right, 0° upward, and 90° clockwise; describe its overall appearance—including shape, color, and texture—and specify any objects or features that would be absent, hidden, or not included from this viewpoint.

BLIP3-o



+



*The plush toy in the image is currently oriented upright with its front facing forward. If we rotate it 0° to the right, 0° upward, and 90° clockwise, the toy will appear as if it has been turned to face its left side. Here's a description of its appearance from this new perspective: - **Shape:** The plush toy maintains its overall rounded, soft, and cuddly shape typical of stuffed toys. - **Color:** The primary colors remain pink and white, with accents of purple on the eyes and some decorative elements like flowers and hearts. - **Texture:** The texture appears soft and plush, consistent with the material typically used for such toys. - **Features Absent, Hidden, or Not Included:** - The front-facing features (like the face, eyes, and mouth) will no longer be visible. - The back of the toy, including any details or patterns on the back, will be hidden from view. - The arms and legs, which are positioned outward in the current orientation, will now be positioned closer together and may appear more compact due to the rotation. In summary, the plush toy will look like a side profile with its left side facing forward, showing less of its front features and more of its side profile.*

実験結果：アブレーション分析








































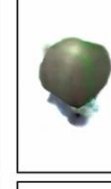










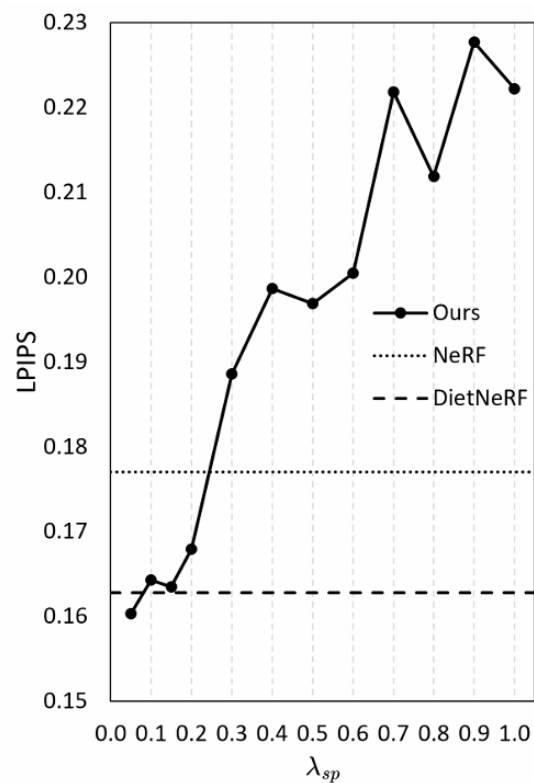
		Image pose range for training									
	GT (Similar pose)										
(a)	DietNeRF										
(b)	$\mathcal{L}_{SC} + \mathcal{L}_{SP}$ + \mathcal{L}_{MSE}										
(c)	$\mathcal{L}_{SP} + \mathcal{L}_{MSE}$										
(d)	$\mathcal{L}_{SC} + \mathcal{L}_{SP}$ + \mathcal{L}_{MSE} (Changed in the prompt)										

図 5.6: アブレーション分析. (a) DietNeRF, (b) 提案手法, (c) 意味予測損失のみを用いた場合, (d) 提案手法でプロンプトを「パンダ」に変えた場合.

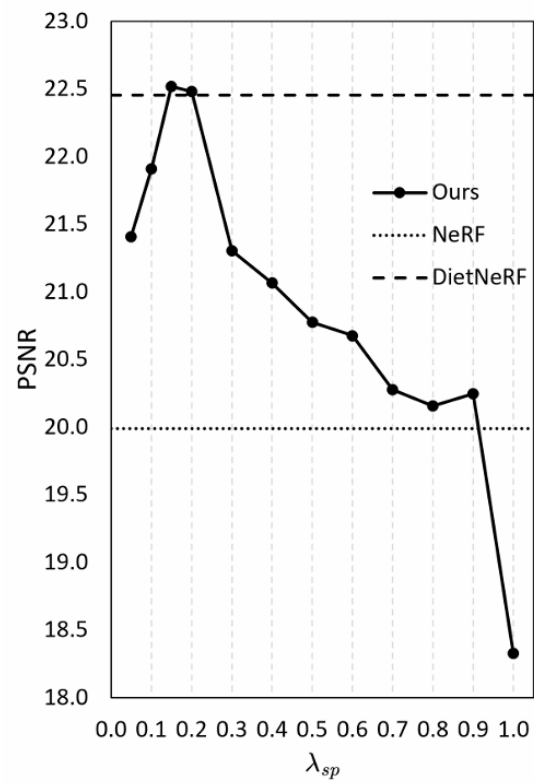
実験結果：アブレーション分析

損失の重みによる変化

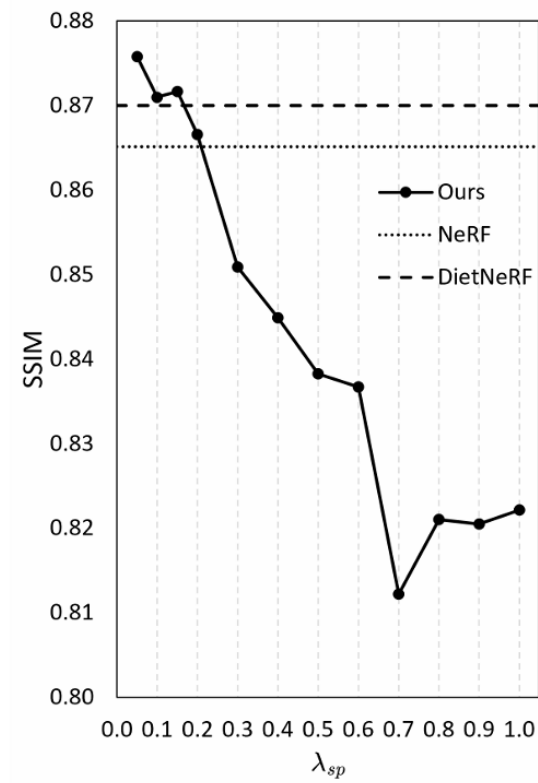
- ・ DietNeRFの意味一貫性損失の重みが0.1と設定されているように、意味予測損失も、0.1以下の値で補助的に導入することで、効果が期待できる



(c) λ_{sp} の変化による LPIPS の推移











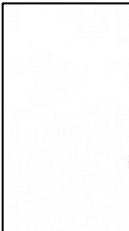
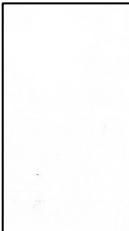
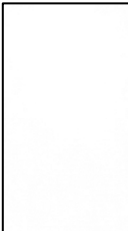


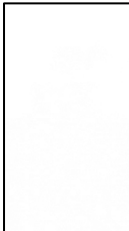

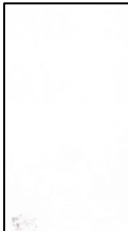
























(a) λ_{sp} の変化による PSNR の推移



(b) λ_{sp} の変化による SSIM の推移

実験結果：アブレーション分析（効果が薄い例）

			Image pose range for training							
GT (Similar pose)										
NeRF										
DietNeRF										
OURS										

おわりに

おわりに

まとめ

- ・ MLLMによる不可視視点の予測を用いたNeRF再構成フレームワークを提案
- ① 未観測領域の意味特徴を取得するための**視点条件付きプロンプト**を導入
- ② **意味予測損失**を導入し学習のガイドとすることで、アーティファクトを抑え、NeRF再構成の**知覚品質を改善**

今後の課題

- ・ 色 / テクスチャ忠実性の強化。拡散モデル併用等による、細かいテクスチャや色一貫性の改善