Appendix

A. Details of GlueNet

The key contribution of this paper is GlueNet, which addresses both cross-model alignment and feature protection. In this section, we will provide more details about GlueNet, including its implementation and configuration, an analysis of its sizes, and other ablation studies.

A.1. Architecture and Configuration

In Section 4.2 of the main paper, we introduced the basic architecture of GlueNet. Different sizes of GlueNet are available, and their configurations are detailed in Table 4. The GlueNet with three residual modules is the smaller variant, with 34M parameters in its encoder, while the GlueNet-5RMs has 51M parameters. However, larger models tend to have slower speed and higher computation costs. Moreover, there could be slight difference in the architecture for different encoders alignment. To maintain inference speed and efficiency during the finetuning stage, the size of GlueNet, when working as an injected module, should not exceed that of the image decoder and text encoder. Based on our empirical observations, assigning RM as five produces satisfactory results. GlueNet-3RMs, on the other hand, appears weak in representation learning. Besides the MLP-mixer, a self-attention-based model may also be used to implement GlueNet. In the future, we plan to explore more suitable architectures for GlueNet.

Table 4. Model configurations for our GlueNet. We introduce two configurations of GlueNet-3RMs and GlueNet-5RMs. [LN] represent that layer-normalization is optional in the Tail Net. DIM-OUT and TOKEN-OUT are 77 and 1024 for Stable Diffusion v1. DIM-IN and TOKEN-IN depend on the target encoder to replace. N is assigned as 1 if TOKEN-IN is equal to TOKEN-OUT. N is larger than 1 (2 or 3) if TOKEN-IN is not equal to TOKEN-OUT.

Stage	Dimensions	Block	GlueNet-3RMs	GlueNet-5RMs
Head Net	TOKEN-IN → TOKEN-OUT	Token MLP	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times N$	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times N$
	DIM-IN →DIM-OUT	Sequence MLP	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times N$	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times N$
Body Net	TOKEN-OUT → TOKEN-OUT	Token MLP	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times 3$	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times 5$
	DIM-OUT →DIM-OUT	Sequence MLP	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times 3$	$\begin{bmatrix}\text{Linear, LN}, \sigma \\ \text{Linear, LN}, \sigma \\ \text{Linear, LN}\end{bmatrix} \times 5$
Tail Net	TOKEN-OUT → TOKEN-OUT	Token MLP	$\begin{bmatrix}\text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}\end{bmatrix}$	$\begin{bmatrix}\text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}\end{bmatrix}$
	DIM-OUT →DIM-OUT	Sequence MLP	$\begin{bmatrix}\text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}\end{bmatrix}$	$\begin{bmatrix}\text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}, \sigma \\ \text{Linear, [LN]}\end{bmatrix}$

B. Analysis of Text Encoder Replacement

B.1. Analysis of GlueNet Sizes

Figure 13 presents a visual comparison of example prompts across three different methods. The original LDM model with the checkpoint [6] shared in its repository is referred to as LDM Ori. Our models, GlueNet-3RMs and GlueNet-5RMs, contain three or five residual modules within their body nets, respectively. The text encoder is replaced with T5-3B, while the image decoder is the same as LDM Ori. Both GlueNet models are trained on the same text corpus, consisting of 18 million English sentences. The figure shows that GlueNet-3RMs struggles with some complex prompts, such as "a virus monster is playing guitar, oil on canvas" and "there is a penguin with a dog head standing," where GlueNet-5RMs performs better. We can, therefore, conclude that deep GlueNet is necessary for precise alignment. GlueNet-5RMs is the default configuration mostly.

[6] https://github.com/CompVis/latent-diffusion

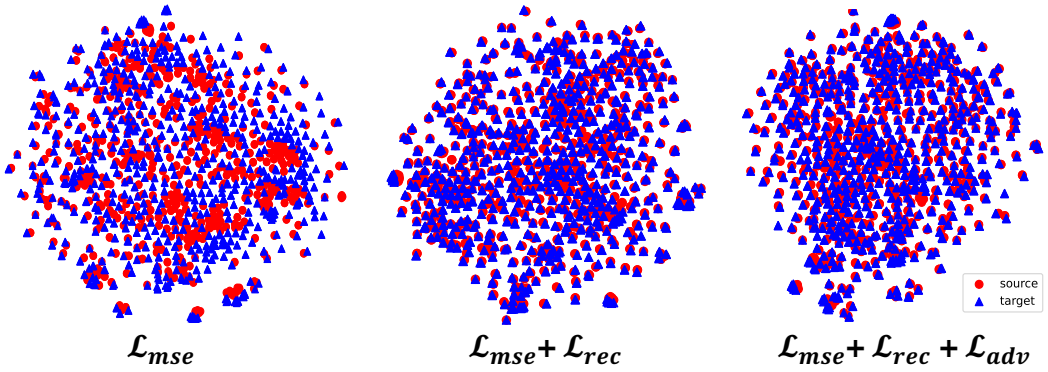

Figure 12. t-SNE [64] cross-model feature map visualization.

Table 5. Ablation study of GlueNet (T5+GlueNet+LDMUnet) over the 5K COCO subset. Finetuning is not applied in most of the cases.

Ablations				s=1.5		s=5		s=7.5	
loss-mse	loss-rec	loss-adv	Finetuning	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓
✗	✗	✗	✗	10.70	98.17	10.75	141.19	11.00	156.70
✗	✗	✗	✓	18.98	35.76	22.05	49.14	22.49	53.11
✓	✗	✗	✗	20.40	34.41	23.01	48.67	23.40	52.78
✓	✓	✗	✗	20.53	33.14	23.23	45.96	23.57	48.67
✓	✓	✓	✗	20.67	32.80	23.24	45.48	23.74	48.51
✓	✓	✓	✓	**21.14**	**30.93**	**23.88**	**41.92**	**24.17**	**44.58**

B.2. Ablation of Losses

In order to analyze the impacts of different losses, we present their t-SNE map visualizations in Figure 12. It is evident from the figure that using just the MSE loss is insufficient to align the features of the two models accurately. The reconstruction loss is crucial to maintain discrimination and avoid overfitting. In contrast, the adversarial loss does not seem to provide significant improvement in the figure. Through empirical study, we found that the adversarial loss only yields limited gains.

We conducted a quantitative ablation study on a randomly selected subset of 5,000 images from COCO dataset. For alignment, we used T5-3B as the text encoder and aligned it with LDM Unet using our GlueNet. The finetuning of Unet is denoted by FT (Unet finetuning is only needed here). We report FID and CLIP scores in the following table for comparing image quality and image-text alignment. The testing data is inferred by DDIM with 200 steps, and the image size is 256 $\times$ 256. For a comprehensive analysis, the experiments were performed using three classifier-free guidances with s=1.5, 5, and 7.5, according to Eq. (2). Table 5 summarizes the results of our ablation study.

The first row of the table represents the direct combination of T5 and LDM, which yields nonsensical results due to severe misalignment. The second row reports results of T5+LDM trained from scratch. The finetuning of Unet in T5+LDM took 100 GPU days, whereas GlueNet training required only 5 GPU days. However, even with ten times the cost, its results were inferior to those of GlueNets'. By comparing the third, fourth, fifth, and sixth rows, we can easily conclude the superiority of the full-version model.

B.3. Text Encoders Analysis

Our proposed framework is compatible with a wide range of text encoders. In this subsection, we experimented with several language models as the text encoder, including Bert-base (110M parameters), Roberta-Large (355M parameters), Clip-text (123M parameters), T5-large (770M parameters), and T5-3B (2.8B parameters). These pre-trained text encoders were aligned with the Latent Diffusion Model (LDM) using our proposed model without requiring any finetuning of the LDM UNet. These plug-and-play models were evaluated on a subset of COCO consisting of 5,000 randomly selected samples. The FID and CLIP scores are reported in the table below for image quality and image-text alignment comparison. The GlueNets were trained on 18 million sentences sampled from the captions of Laion-400M. The testing data were inferred using DDIM with 200 steps, and the image size was set to 256×256. To provide a comprehensive analysis, we conducted experiments on

Table 6. Analysis of Different Text Encoders over the 5K COCO subset.

	Bert-B (110M)		Roberta-L (355M)		CLIP-Text (123M)		T5-L (770M)		T5-3B (2.8B)	
	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	FID↓
s=1.5	19.42	37.46	20.03	34.06	19.02	38.93	19.78	35.17	**20.67**	**32.80**
s=5	21.97	55.35	22.85	45.88	21.85	47.32	22.76	47.91	**23.24**	**45.48**
s=7.5	22.68	55.21	23.25	48.96	22.38	49.66	23.23	50.02	**23.74**	**48.51**

Table 7. Model Transfer with variant token lengths (SrcTokenLength →TargetTokenLength) over the 5K COCO subset with the guidance as 5. Roberta-L [31] is applied as the new text encoder to replace LDM text encoder.

	77→77	128→77	256→77
CLIP↑	22.85	23.19	**23.37**
FID↓	45.88	45.67	**45.53**

Table 8. Analysis of GlueNet's training cost (T5-3B → LDM) with increasing sizes of training data. The FID and CLIP scores are computed over 5K subset of COCO.

	Data Size		
	5M	18M	116M
CLIP↑	20.92	23.24	**23.71**
FID↓	48.63	45.48	**43.17**
GPU Days↓	**1.67**	5.89	41.20

three classifier-free guidance settings with s=1.5, 5 and 7.5. The results are summarized in Table 6. As shown in the table, we observe that both FID and CLIP scores increase with the use of larger text models, which is consistent with the findings reported in Imagen [48]. However, CLIP-text does not perform well in our experiments due to a larger domain gap with text-only models like Bert.

B.4. Variant Token Lengths

GlueNet demonstrates strong ability in handling text of different lengths. The token number is fixed at 77 for both LDM and SDM. Our proposed GlueNet can handle variable length text encoders without any finetuning of the Unet model. To verify this, we conducted an experimental study, and the results are reported in Table 7. In this experiment, we used Roberta-L and its tokenizers to encode text with maximum tokens of 77, 128, and 256, respectively. The guidance setting was 5, and other experimental settings were consistent with those in Table 6.

B.5. Data Sizes

Conditional generation is a challenging task, and the alignment between the text encoder and image Unet remains an open question. We have found empirically that replacing the text encoder can yield comparable results to the source model, but it requires significant effort to surpass it. The bottleneck may lie in the Unet architecture. In this section, we provide precise computations costs for different training set sizes in Table 8. The benchmark was performed on COCO-5K using T5-3B to replace the LDM text encoder with GlueNet, consistent with previous experiments. As shown in the table, we observe that performance increases with both the size of the training set and the computation cost.

C. Monolingual (English) Text-to-Image Generation

To demonstrate the generality of our proposed framework, we also replaced the CLIP text encoder of Stable Diffusion (v1-4) with T5-Large. As shown in Figure 14, our model exhibits precise controllability and excellent visual quality compared to the standard Stable Diffusion model. However, it falls short of outperforming the original Stable Diffusion model due to minor mismatches.

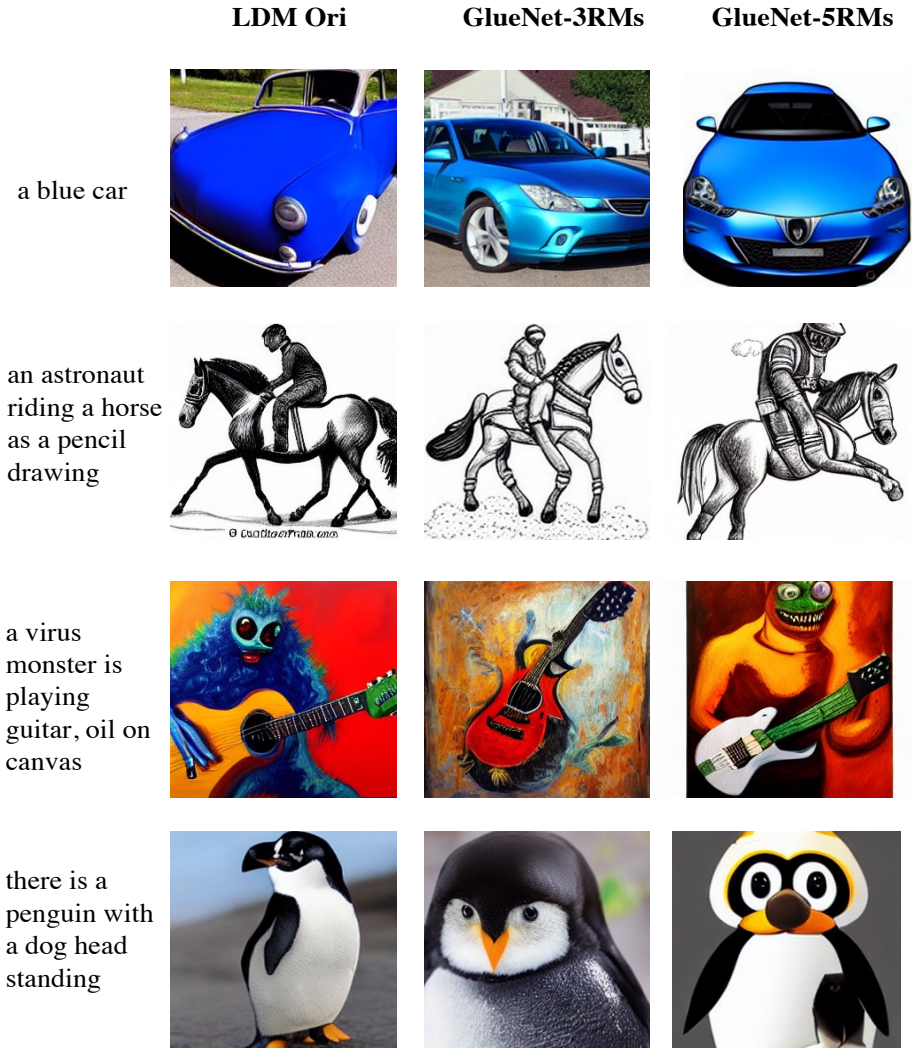

Figure 13. Monolingual generation (T5 + GlueNet + LDMUnet) of example prompts in 256 $\times$ 256 with guidance weight 7.5 and DDIM steps 200. Both T5 and LDMUnet are pre-trained ones. We only train GlueNet with different model sizes to fulfill.

Figure 14. Monolingual generation of example prompts in 512 × 512 with guidance weight 7.5 and 50 PLMS [30] sampling steps.

D. Multilingual Text-to-Image Generation

The additional results in Figures 19, 20, 21, 22 and 23 help to verify the multilingual text-to-image generation capabilities of our proposed framework. Each language is associated with a dedicated GlueNet. To build an automated pipeline, the model should be able to select the appropriate GlueNet based on the detected language.

E. Sound-to-Image Generation

We provide additional visual results in Figures 16 and 17, comparing the vanilla GlueNet (without the re-weight objective) with the Adaptive GluNet (with the re-weight objective) as described in Section 4.3 of the main paper. These figures clearly show that the Adaptive GluNet improves the stability and accuracy of the generated images compared to the vanilla GlueNet, thus demonstrating the effectiveness of the objective re-weighting technique proposed in Section 4.3.

F. Multimodal-to-Image Generation

According to Section 4.3 of the main paper, the fusion of multi-modal condition features from different encoders can be achieved through non-parametric operations, such as concatenating the top K signals and averaging the rest (excluding the last K). More multi-modal generation results are presented in Figure 18. In our experiment, we used a text encoder (CLIPText) to extract the text embedding (with the input of "in painting style by Picasso"). We also inputted audio data to the AudioCLIP model, which was appended with a GlueNet to map it into an embedding. Then, we applied the proposed feature fusion operator, merging both text and sound embeddings, which enabled the stable diffusion model to generate reasonable results.

We have conducted an ablation study to determine the optimal value of K for fusing multi-modal condition features using the non-parametric operations described in Section 4.3 of the main paper. The results, which are presented in Figure 15, show that when $K <= 6$, the audio signals dominate the generation process. As K increases, text signals gradually begin to appear and eventually dominate the conditioning when $K = 10$. Therefore, to strike a good balance between the two cross-modal signals, we recommend selecting an appropriate value of K based on empirical observations.

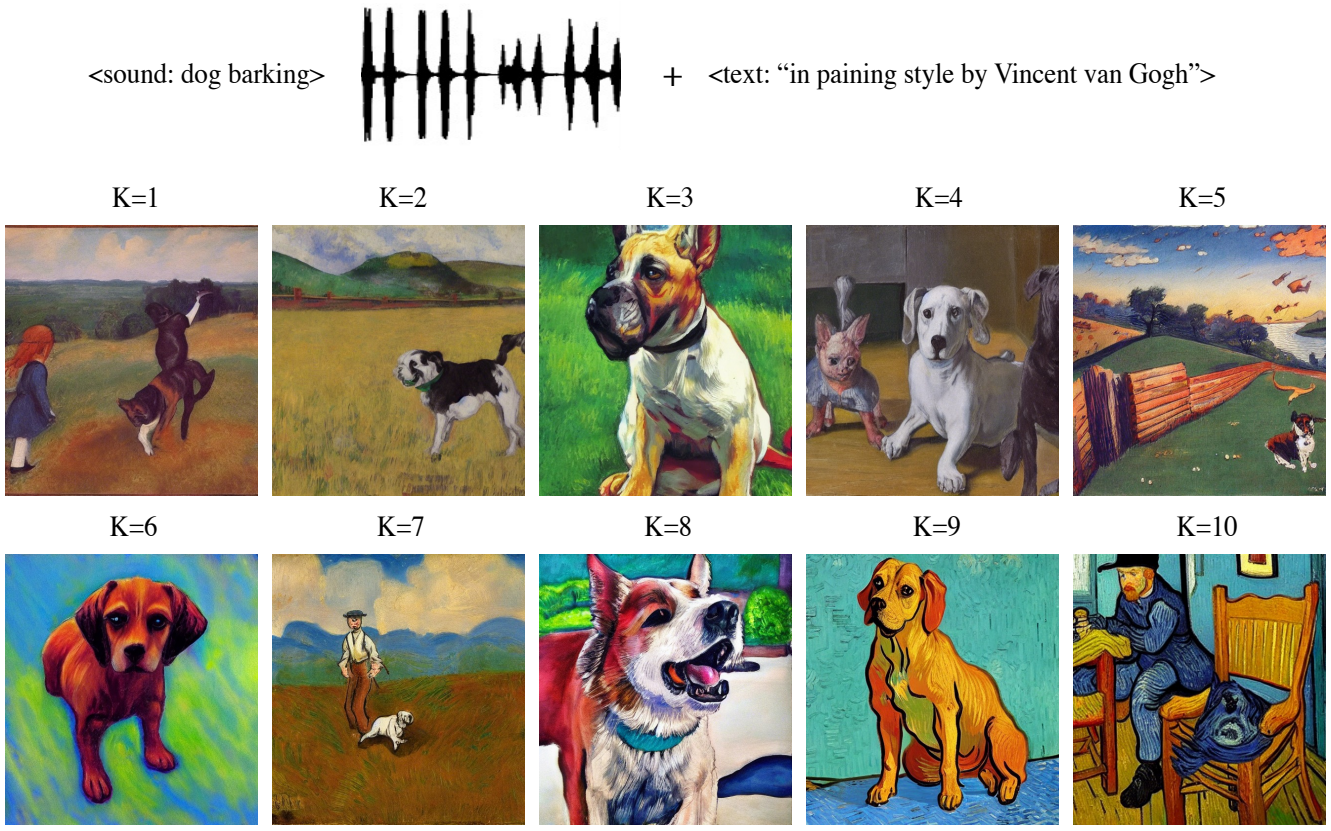

Figure 15. Analysis of top K selected signals for multimodal (sound and text) feature fusion. This operation is non-parametric (also training-free) which only needs concatenating top K token signals and averaging the rest.

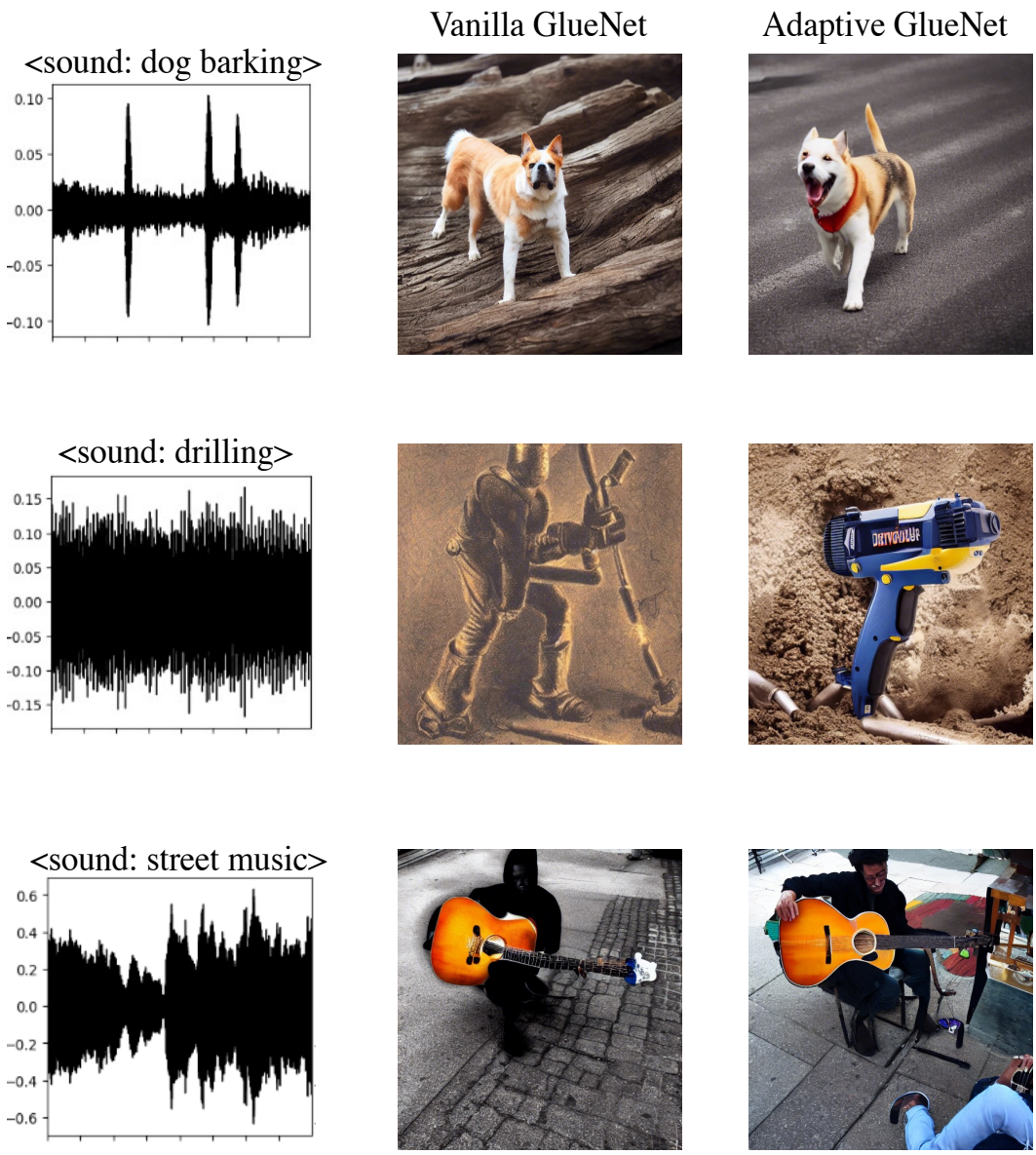

Figure 16. Sound-to-image generation (Part 1/2) with vanilla and adaptive GlueNets.

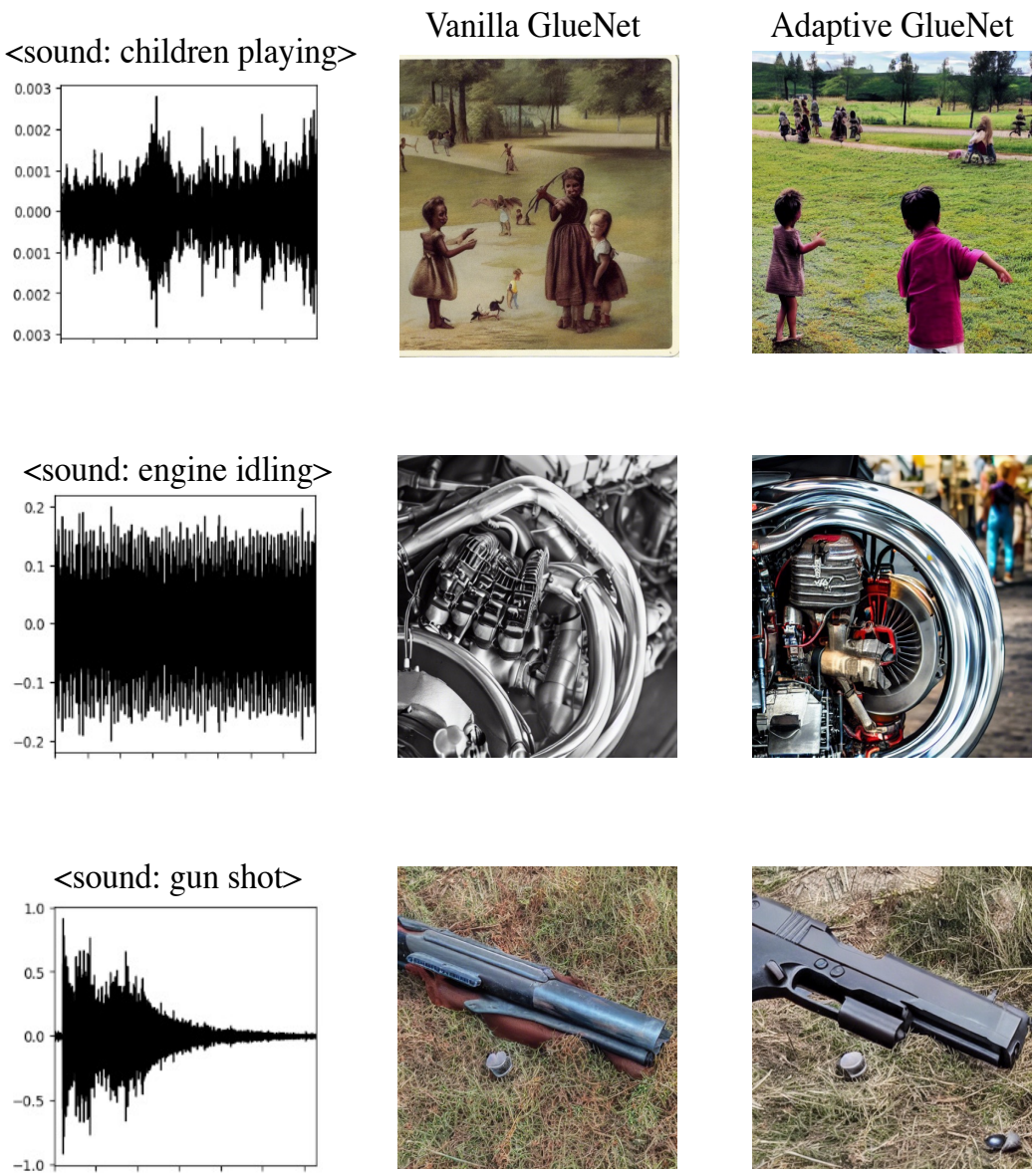

Figure 17. Sound-to-image generation (Part 2/2) with vanilla and adaptive GlueNets.

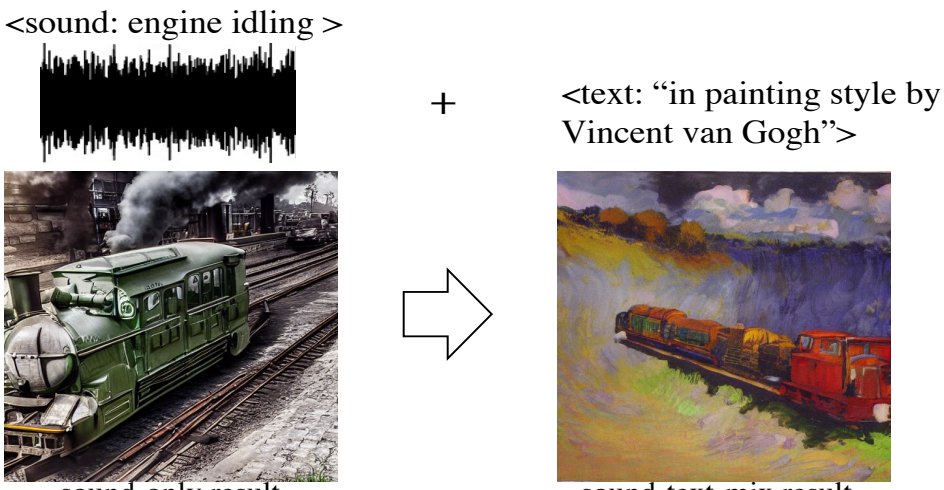

Figure 18. Multimodal-to-image generation (sound-text-mix) results. The condition encoders are ClipText and AudioCLIP+GlueNet. The two-modality features are fused to Stable Diffusion Unet to generate such right-side result.

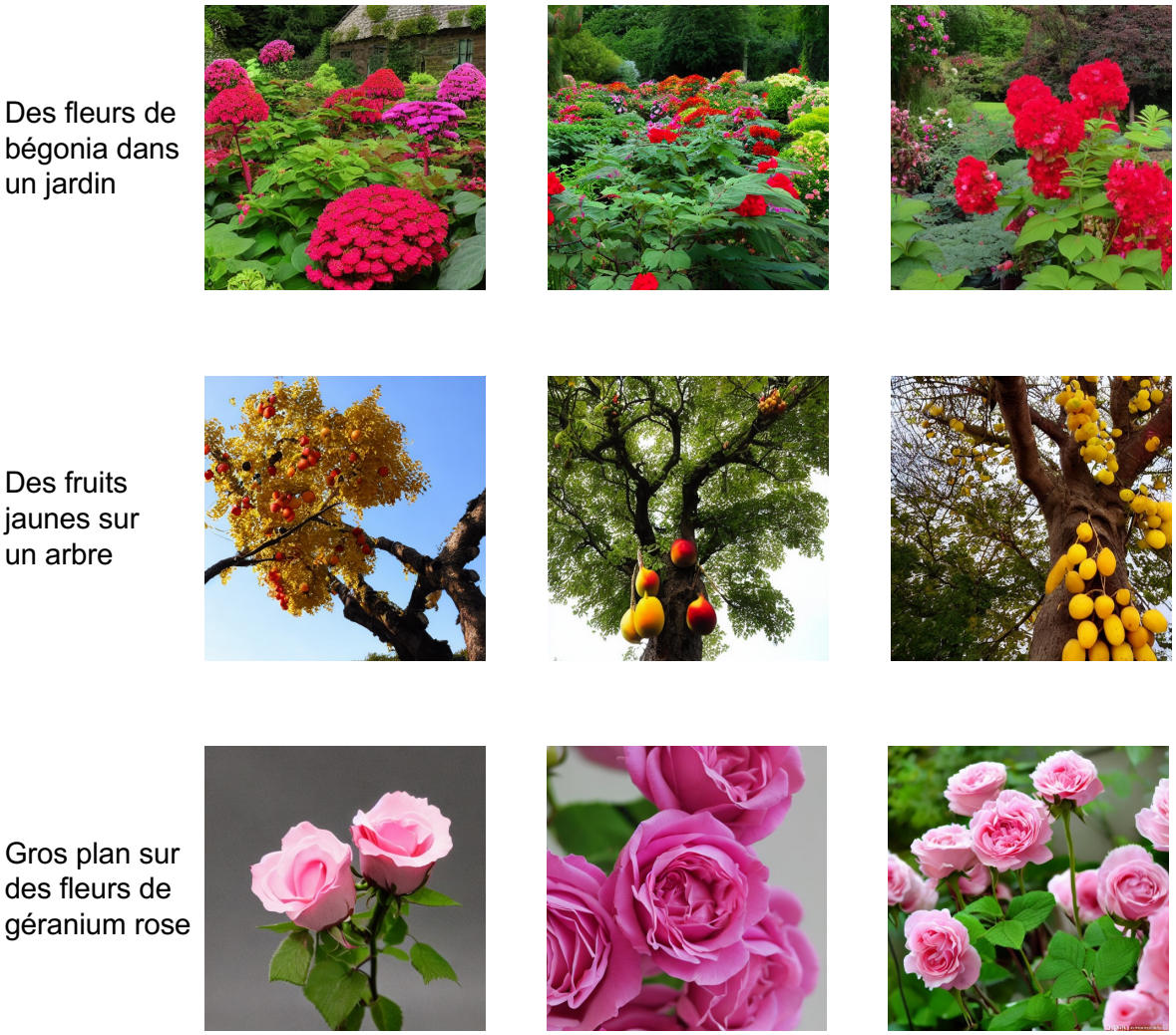

Figure 19. Multilingual generation results in 512 $\times$ 512 of XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) with the French captions. The three results are generated with different random noises.

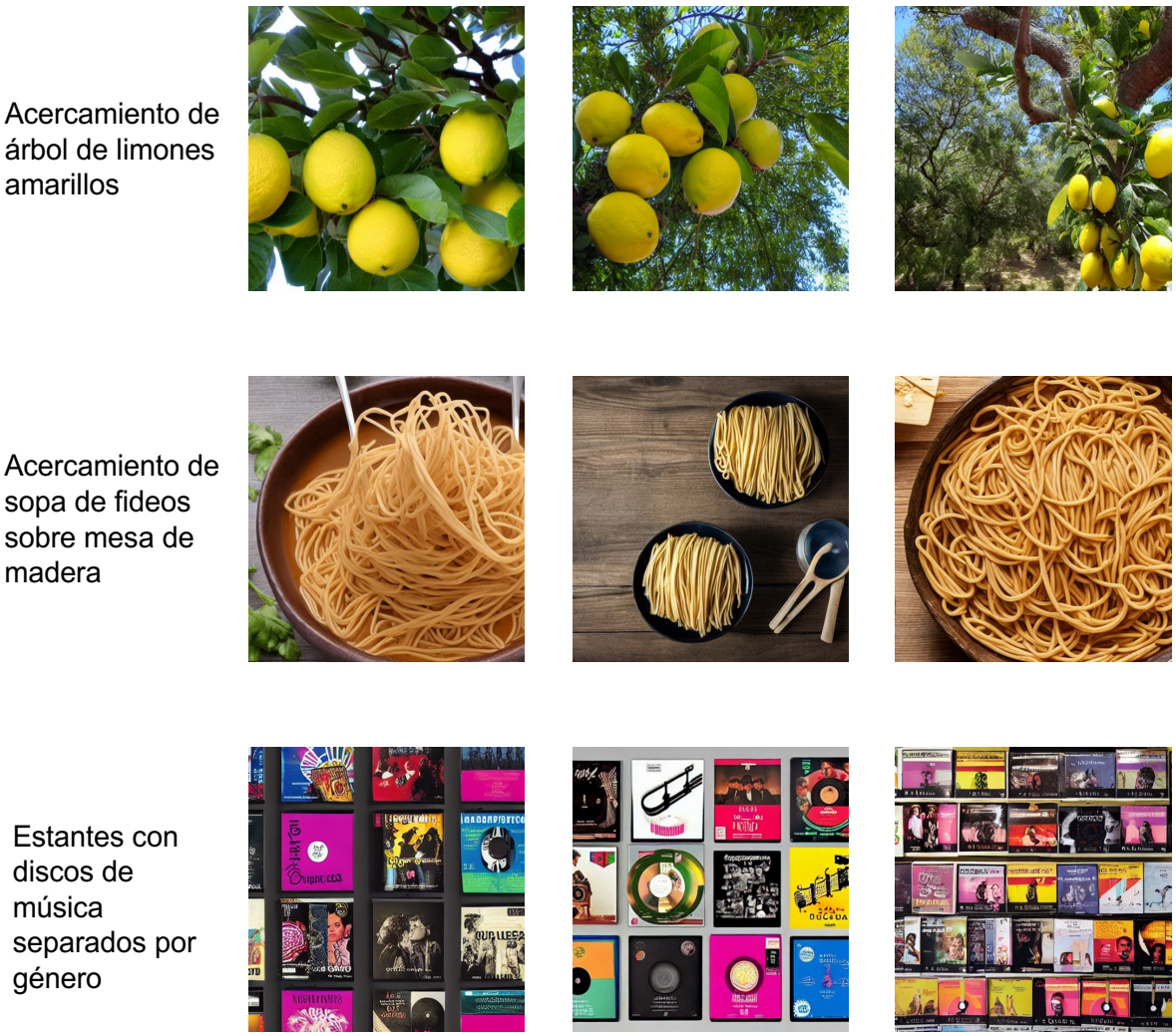

Figure 20. Multilingual generation results in 512 $\times$ 512 of XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) with the Spanish captions. The three results are generated with different random noises.

一块木板上盛放着刚出炉的圆形披萨，上面撒了一些绿叶

两个古董意大利车

在山里中站着两只鸡，一只黄色另一只黑黄色，它们俩站着看向同一个方向

Figure 21. Multilingual generation results in 512 $\times$ 512 of XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) with the Chinese captions. The three results are generated with different random noises.

Figure 22. Multilingual generation results in 512 × 512 of XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) with the Japanese captions. The three results are generated with different random noises.

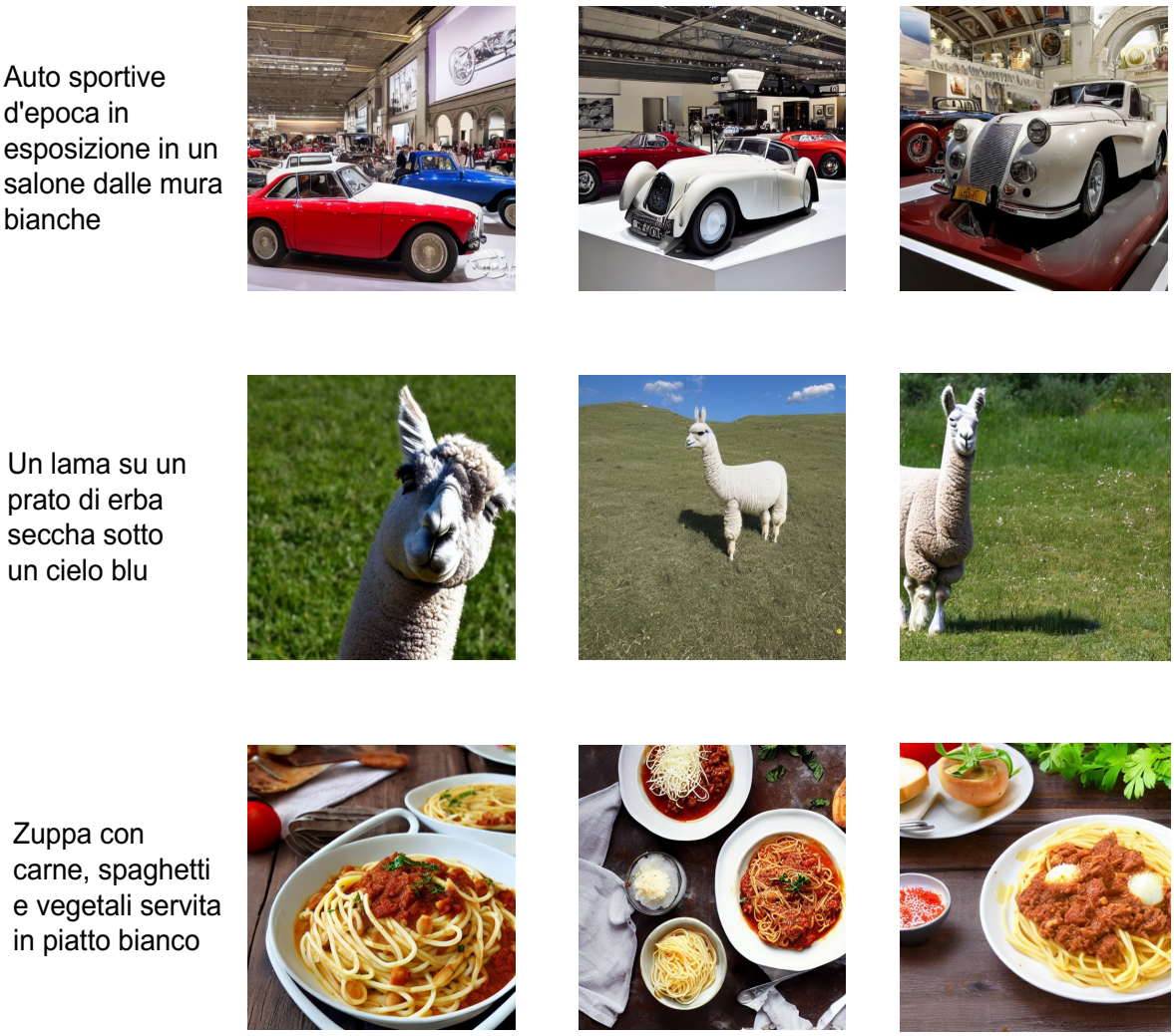

Figure 23. Multilingual generation results in 512 $\times$ 512 of XLM-Roberta + GlueNet + Stable Diffusion Unet (v1-5) with the Italian captions. The three results are generated with different random noises.