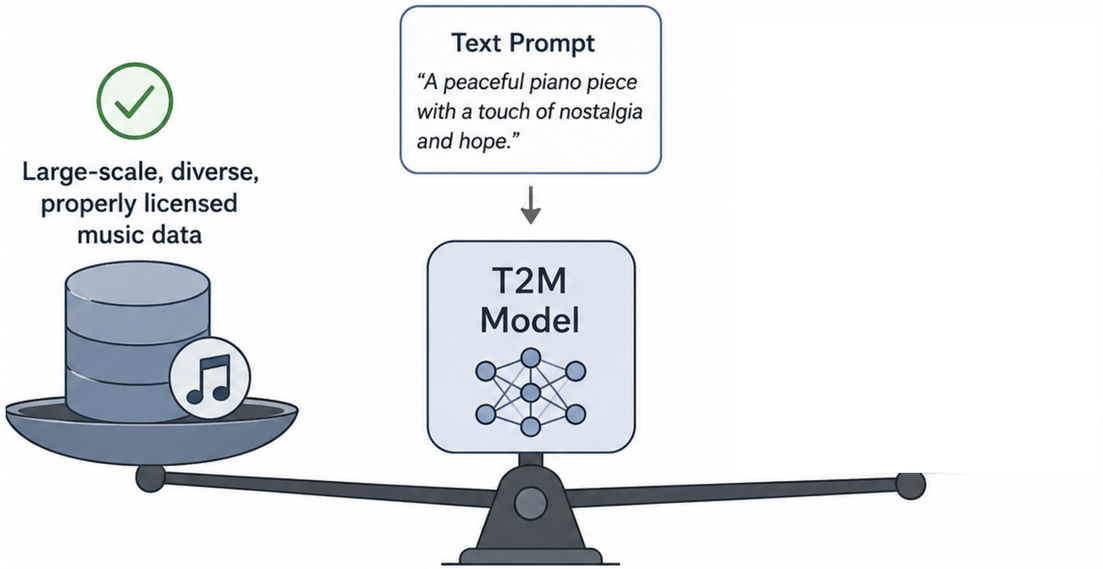
PADS-TAL: Padding-Annealed Diffusion Sampling in Text-Aware Latent Space for Robust and Diverse Text-to-Music Generation

Taekoan Yoo[1], Wonkyung Jung[1], KyungHun Kim[1], Kyeongbo Kong[2]

[1] AI Tech Lab., NHN Corp. , [2] Pusan National University

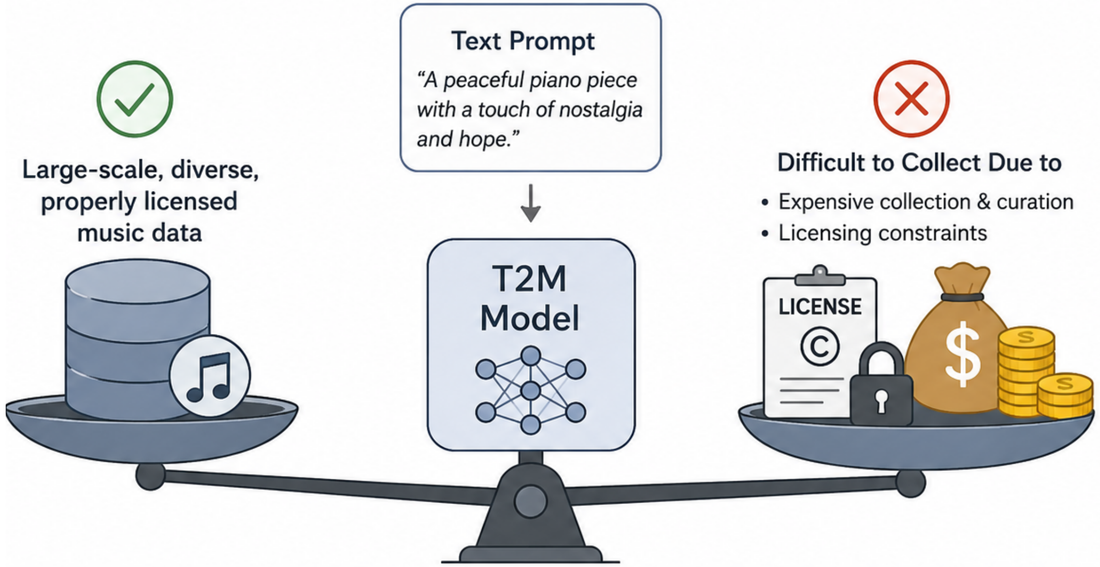
1. Problem

- **Practical Deployment for Text-to-Music**
  - **Requirements**
    - Large-scale,
    - Diverse,
    - and properly licensed music data.



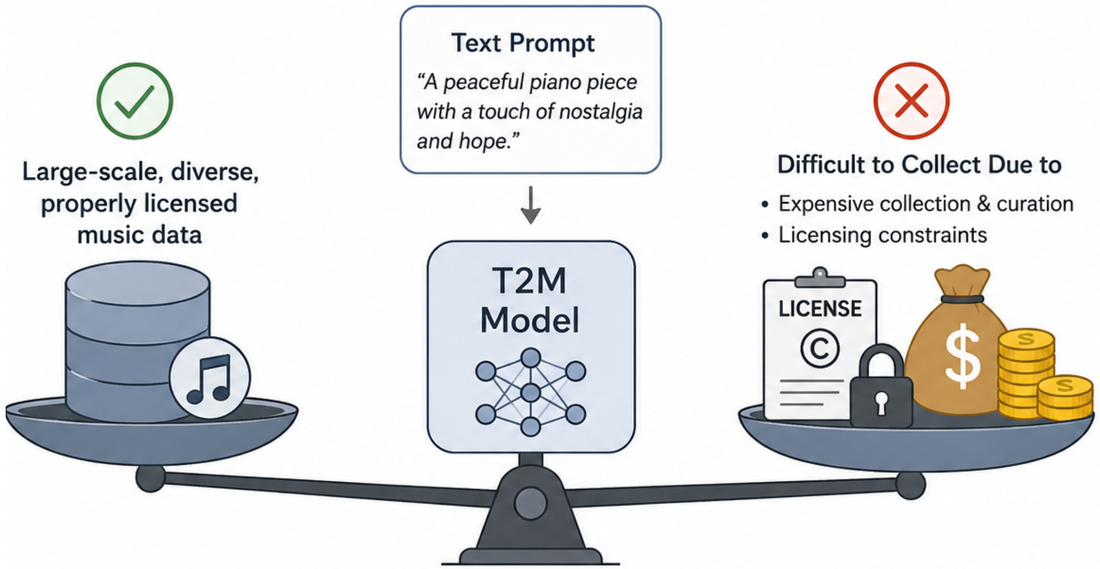
1. Problem

- **Practical Deployment for Text-to-Music**
  - **Requirements**
    - Large-scale,
    - Diverse,
    - and properly licensed music data.
  - **However, collecting such data is difficult**
    - Expensive data collection & curation
    - Licensing constraints
  - **As a result, T2M models may exhibit:**
    - Limited diversity
    - Repetitive generation patterns



1. Problem

- **Practical Deployment for Text-to-Music**
  - **Requirements**
    - Large-scale,
    - Diverse,
    - and properly licensed music data.
  - **However, collecting such data is difficult**
    - Expensive data collection & curation
    - Licensing constraints
  - **As a result, T2M models may exhibit:**
    - Limited diversity
    - Repetitive generation patterns
  - Nevertheless,
    - methods for improving this form of diversity remain **relatively underdeveloped.**



2.1. Padding Annealed Diffusion Sampling

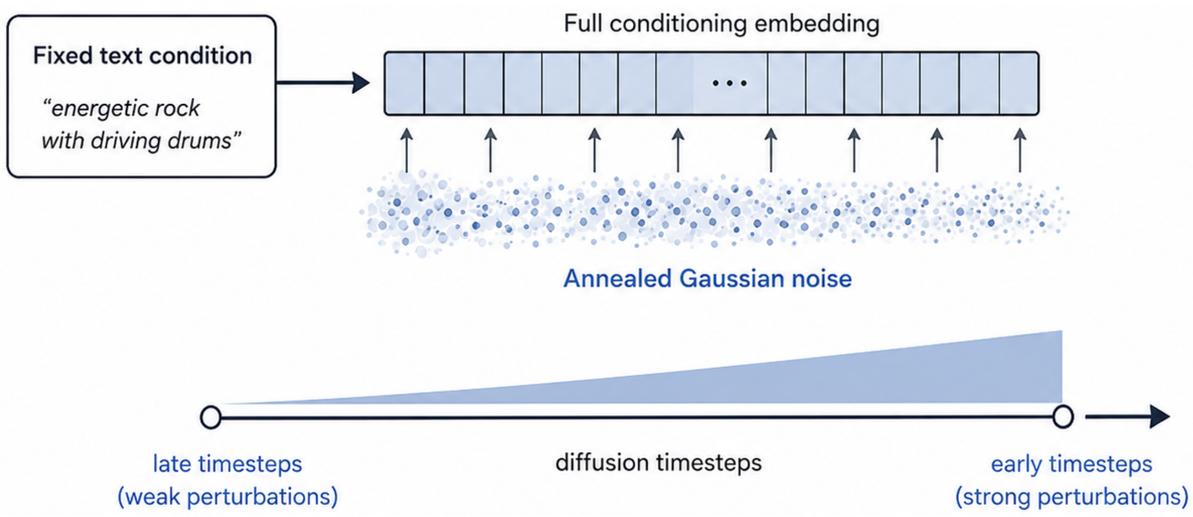
- **How Can We Improve Diversity in Diffusion Models?**
  - **In the domain of image generation,**
    - **numerous inference-time techniques**
    - e.g., CADS, SPARKE, Particle Guidance, ...

2.1. Padding Annealed Diffusion Sampling

- **How Can We Improve Diversity in Diffusion Models?**
  - **In the domain of image generation,**
    - **numerous inference-time techniques**
    - e.g., CADS, SPARKE, Particle Guidance, ...
  - **“CADS” is a natural starting point for T2M**
    - injects annealed Gaussian noise
    - into the full conditioning embedding

$$\hat{\boldsymbol{c}} = \sqrt{\gamma(t)}\,\boldsymbol{c} + s\sqrt{1-\gamma(t)}\,\boldsymbol{n},$$

$$\textbf{where } \gamma(t) = \begin{cases} 1 & \textbf{if } 0 \le t \le \tau_1, \\ \dfrac{\tau_2 - t}{\tau_2 - \tau_1} & \textbf{if } \tau_1 < t < \tau_2, \\ 0 & \textbf{if } \tau_2 \le t \le 1. \end{cases}$$



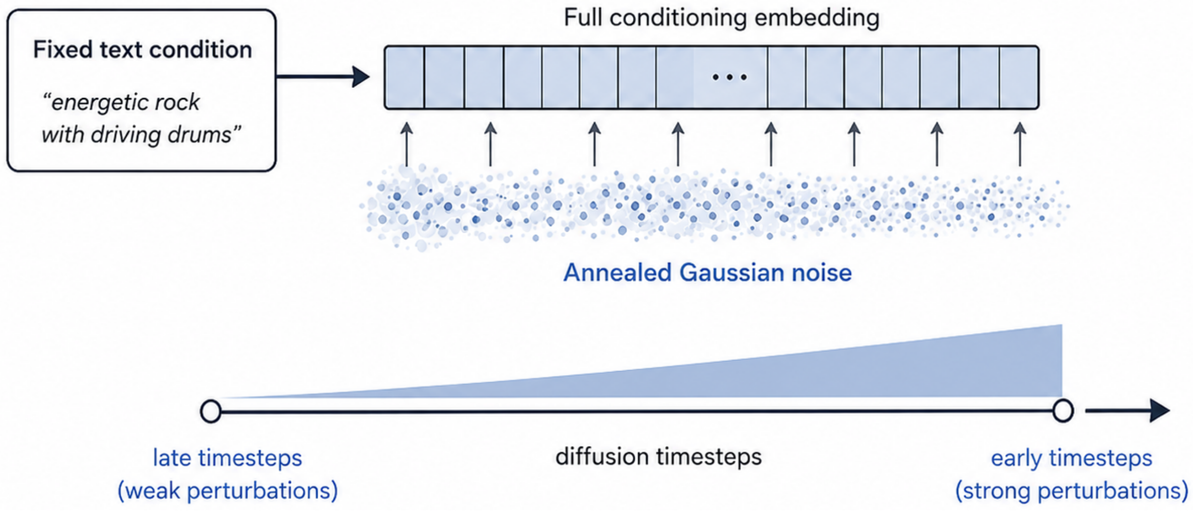
2.1. Padding Annealed Diffusion Sampling

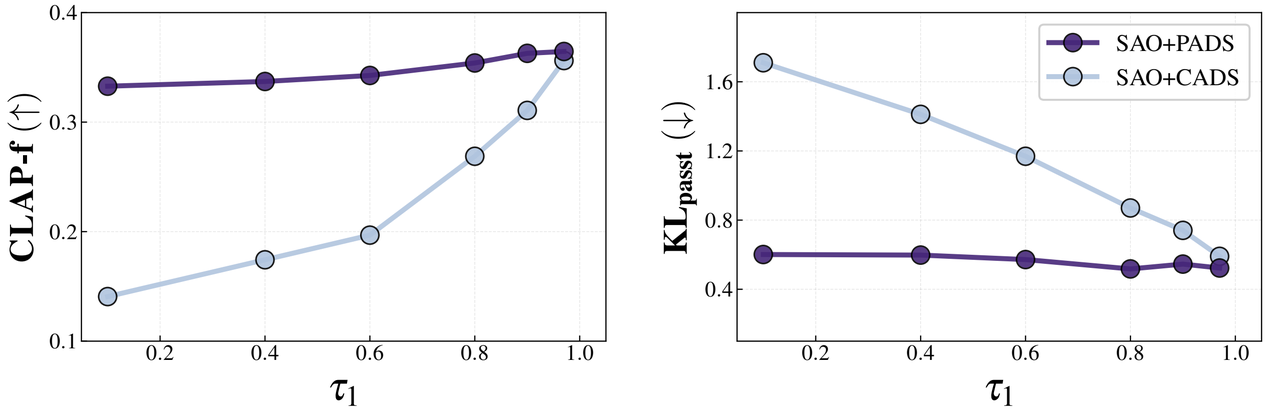
- **How Can We Improve Diversity in Diffusion Models?**
  - **In the domain of image generation,**
    - **numerous inference-time techniques**
    - e.g., CADS, SPARKE, Particle Guidance, ...
  - **“CADS” is a natural starting point for T2M**
    - injects annealed Gaussian noise
    - into the full conditioning embedding

$$\hat{\boldsymbol{c}} = \sqrt{\gamma(t)}\,\boldsymbol{c} + s\sqrt{1-\gamma(t)}\,\boldsymbol{n},$$

$$\textbf{where } \gamma(t) = \begin{cases} 1 & \textbf{if } 0 \le t \le \tau_1, \\ \dfrac{\tau_2 - t}{\tau_2 - \tau_1} & \textbf{if } \tau_1 < t < \tau_2, \\ 0 & \textbf{if } \tau_2 \le t \le 1. \end{cases}$$

  - **However, naive transfer to T2M leads to:**
    - Sharp drop in text alignment
    - Sharp drop in audio fidelity
    - Even with high guidance ω_{cfg}





Text alignment (left) and fidelity (right) as τ_1 varies

2.1. Padding Annealed Diffusion Sampling

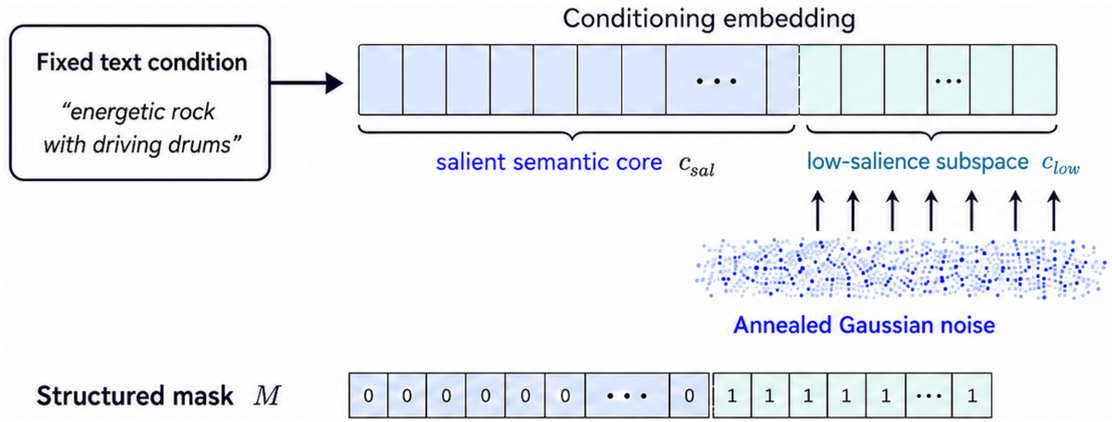
- **PADS: Mask-Controlled Conditioning Perturbation**
  - **Key idea:** perturb only the low-salience part of the conditioning

$$\boldsymbol{c} = [\boldsymbol{c}_{\text{sal}}; \boldsymbol{c}_{\text{low}}]$$

    - with the selected low-salience subspace mask

$$\hat{\boldsymbol{c}} = \boldsymbol{c} \odot (1 - M) + \left(\sqrt{\gamma(t)}\, \boldsymbol{c} + s\sqrt{1 - \gamma(t)}\, \boldsymbol{n} \right) \odot M$$

$$M_i = \begin{cases} \mathbf{1}_D, & \textbf{if } i > L_{\text{sal}} \textbf{ or } i > L - L_{\min}, \\ \mathbf{0}_D, & \textbf{otherwise}. \end{cases}$$



2.1. Padding Annealed Diffusion Sampling

- **PADS: Mask-Controlled Conditioning Perturbation**
  - **Key idea:** perturb only the low-salience part of the conditioning

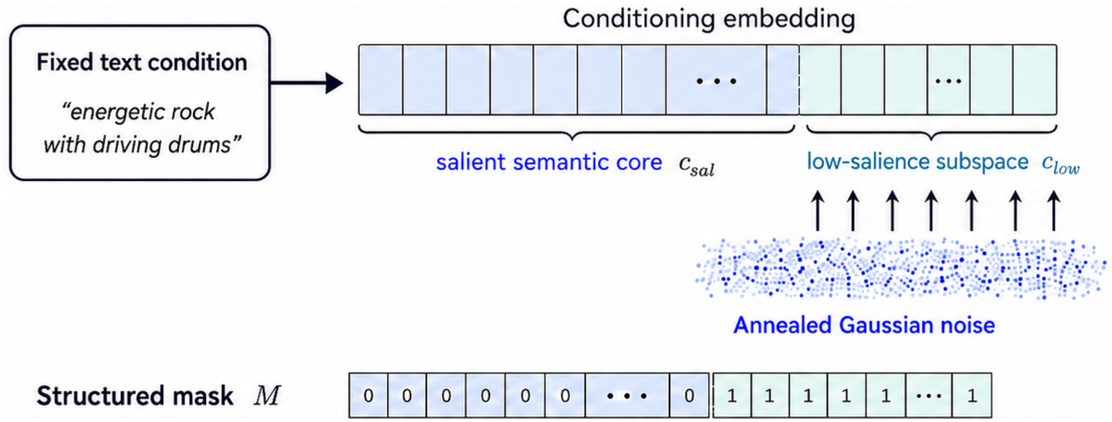
$$\boldsymbol{c} = [\boldsymbol{c}_{\text{sal}}; \boldsymbol{c}_{\text{low}}]$$

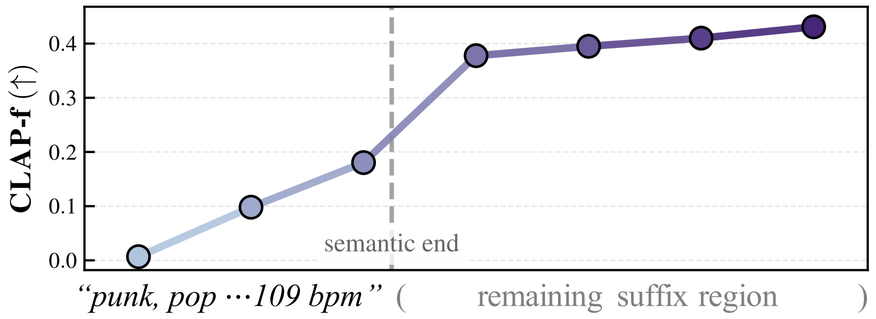
    - with the selected low-salience subspace mask

$$\hat{\boldsymbol{c}} = \boldsymbol{c} \odot (1 - M) + \left(\sqrt{\gamma(t)}\,\boldsymbol{c} + s\sqrt{1-\gamma(t)}\,\boldsymbol{n}\right) \odot M$$

$$M_i = \begin{cases} \mathbf{1}_D, & \text{if } i > L_{\text{sal}} \text{ or } i > L - L_{\min}, \\ \mathbf{0}_D, & \text{otherwise}. \end{cases}$$

  - **Instantiating $\boldsymbol{c}_{\text{low}}$ in T2M : with empirical support**
    - **padding-to-semantic noise sweep**
    - perturbation reaches semantic tokens → alignment drops sharply
    - instantiates $\boldsymbol{c}_{\text{low}}$ as the padding-indexed subspace





padding-to-semantic noise sweep

2.1. Padding Annealed Diffusion Sampling

- **PADS: Mask-Controlled Conditioning Perturbation**
  - **Key idea:** perturb only the low-salience part of the conditioning

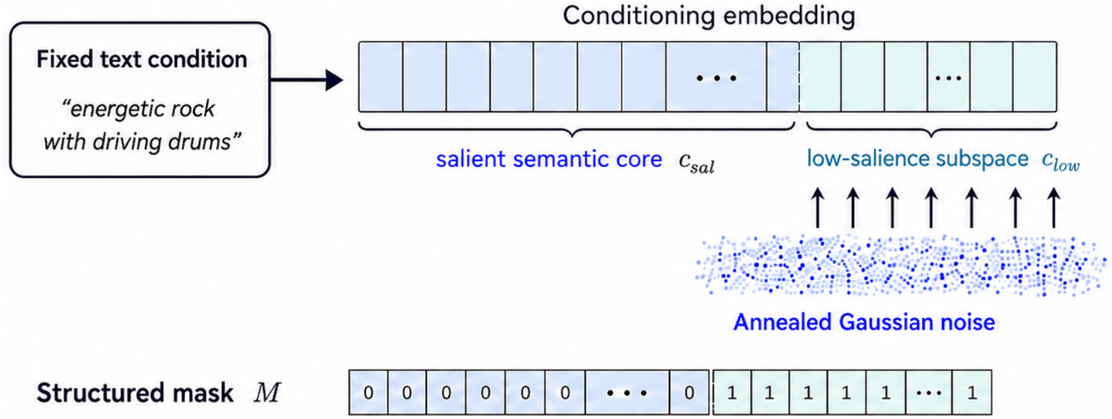
$$\boldsymbol{c} = [\boldsymbol{c}_{\text{sal}}; \boldsymbol{c}_{\text{low}}]$$

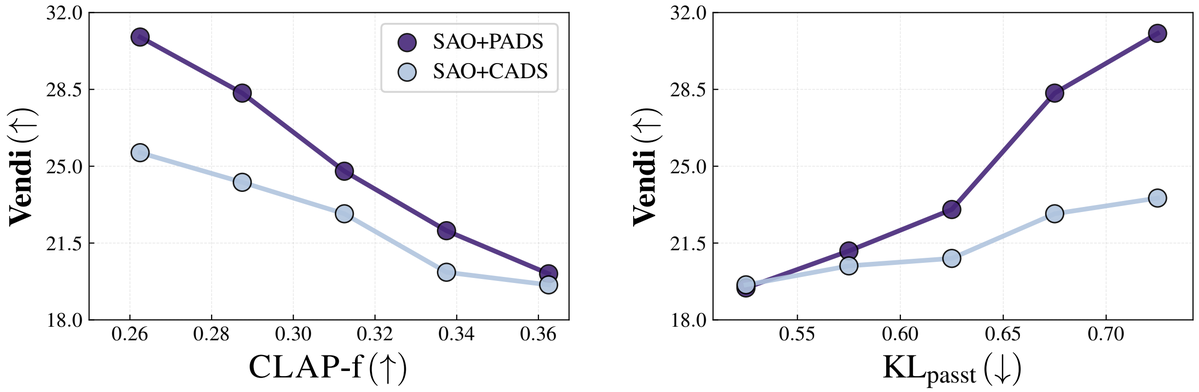
    - with the selected low-salience subspace mask

$$\hat{\boldsymbol{c}} = \boldsymbol{c} \odot (1 - M) + \left(\sqrt{\gamma(t)}\,\boldsymbol{c} + s\sqrt{1-\gamma(t)}\,\boldsymbol{n}\right) \odot M$$

$$M_i = \begin{cases} \mathbf{1}_D, & \textbf{if } i > L_{\text{sal}} \textbf{ or } i > L - L_{\min}, \\ \mathbf{0}_D, & \textbf{otherwise}. \end{cases}$$

  - **Instantiating $\boldsymbol{c}_{\text{low}}$ in T2M : with empirical support**
    - **padding-to-semantic noise sweep**
    - perturbation reaches semantic tokens → alignment drops sharply
    - instantiates $\boldsymbol{c}_{\text{low}}$ as the padding-indexed subspace
  - **As a result, Matched comparison with CADS**
    - better diversity–alignment trade-off
    - better diversity–fidelity trade-off

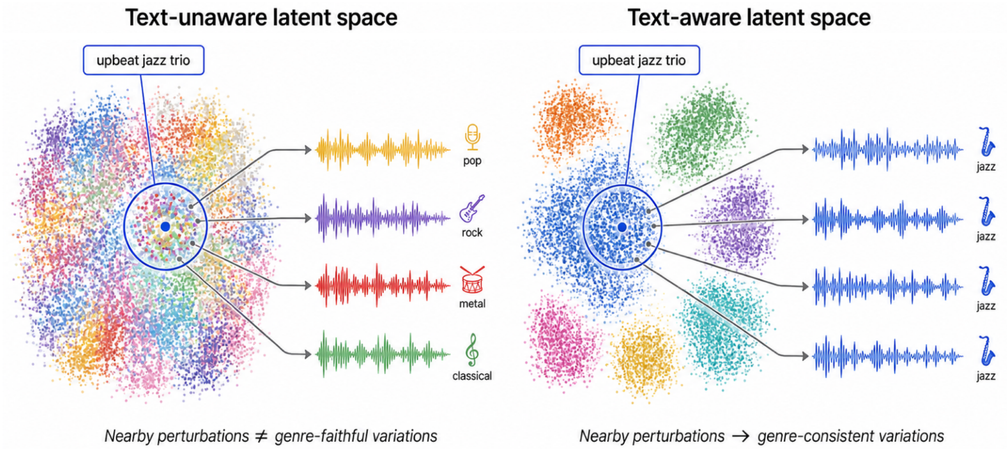




Diversity at matched Diversity(left) or Fidelity(right)

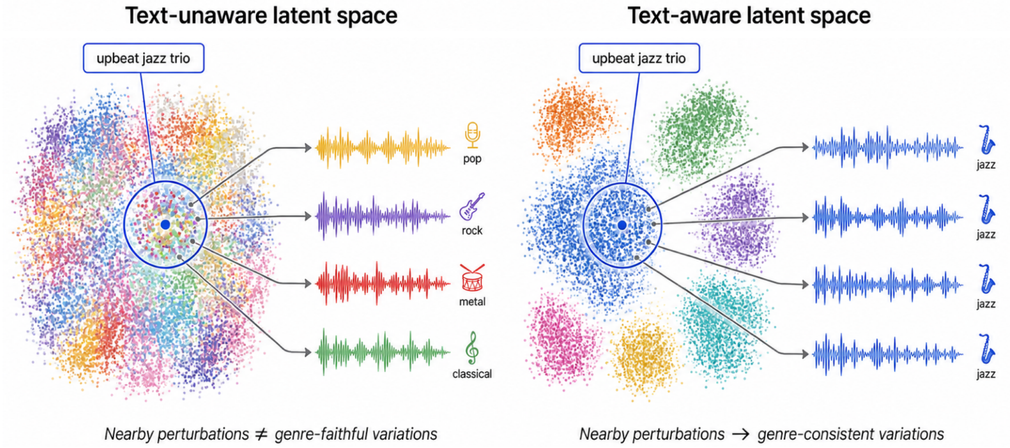
2.2. Text-Aware Latent Space

- **Remaining Issue**
  - **Unstable genre-consistent diversity**
  - **Standard LDM = text-unaware latent space**
    - audio-only VAE latent
      → nearby samples ≠ similar-genre variations
      → not necessarily aligned with intended semantics
    - conditional DM sampling in that space
      → little guarantee of text-semantic neighborhood structure

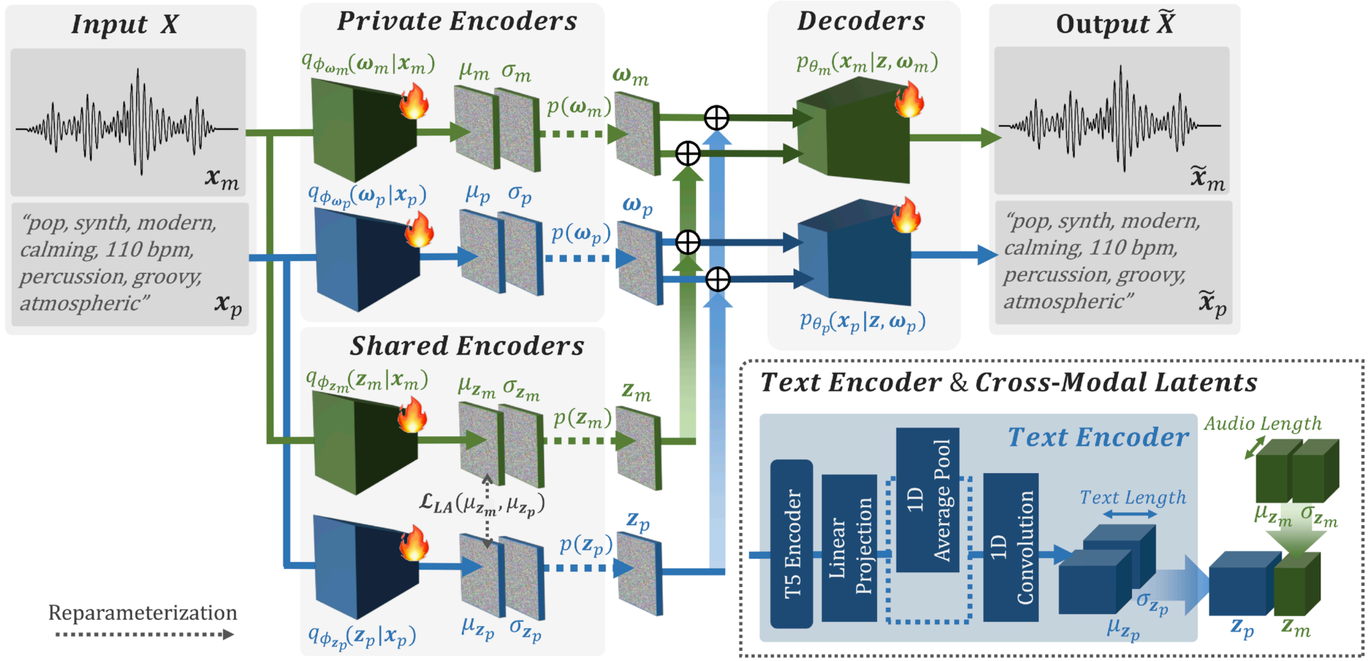


2.2. Text-Aware Latent Space

- **Remaining Issue**
  - **Unstable genre-consistent diversity**
  - **Standard LDM = text-unaware latent space**
    - audio-only VAE latent
      → nearby samples ≠ similar-genre variations
      → not necessarily aligned with intended semantics
    - conditional DM sampling in that space
      → little guarantee of text-semantic neighborhood structure

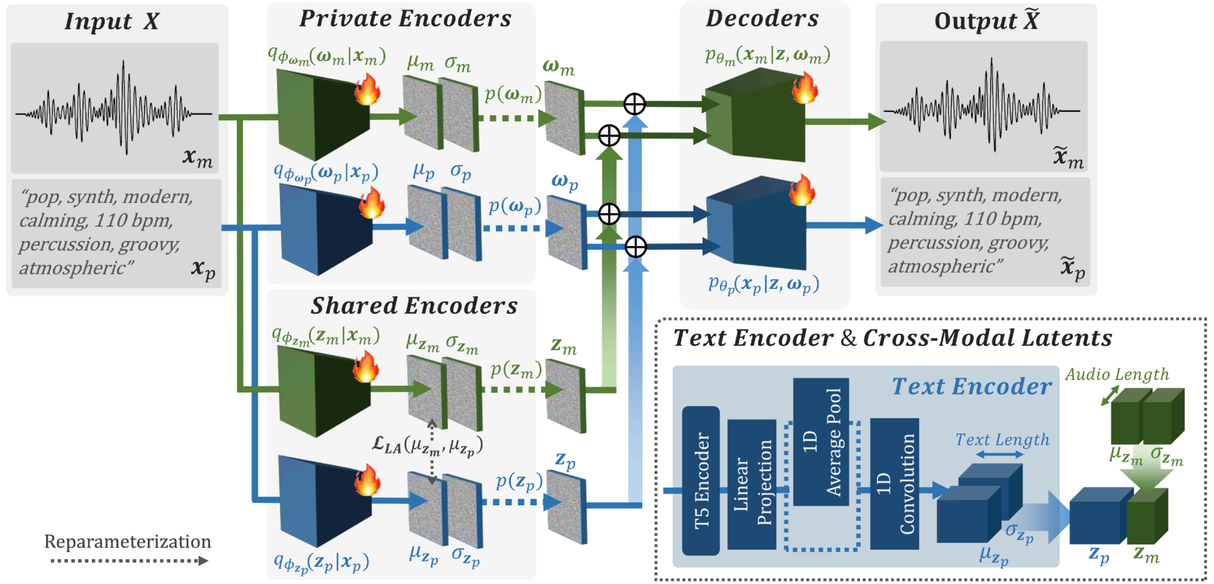


- **Training a VAE for the TAL Space**
  - **Based on prior MoE-style multimodal VAE**
    - shared–private latent factorization
  - **Two types of encoders**
    - private latents → modality-specific details
    - shared latents → text–audio semantics



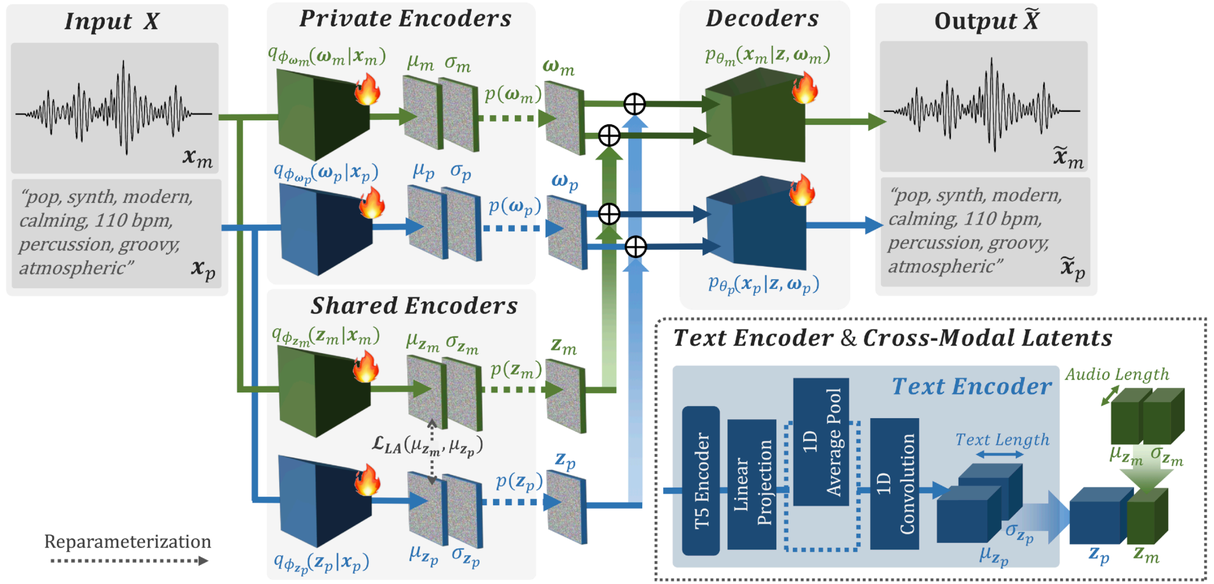
2.2. Text-Aware Latent Space

- **MoE-style mVAE Objective for the TAL Space**
  - **Stable Audio Open: audio-only VAE training**
    - $\mathcal{L}_{VAE} = \alpha_{adv}\mathcal{L}_{adv} + \alpha_{mrstft}\mathcal{L}_{mrstft} + \alpha_{kl}\mathcal{L}_{kl}$
      - ** $\mathcal{L}_{adv}$: an adversarial loss
      - ** $\mathcal{L}_{mrstft}$: an audio reconstruction loss
      - ** $\mathcal{L}_{kl}$: a KL divergence term



2.2. Text-Aware Latent Space

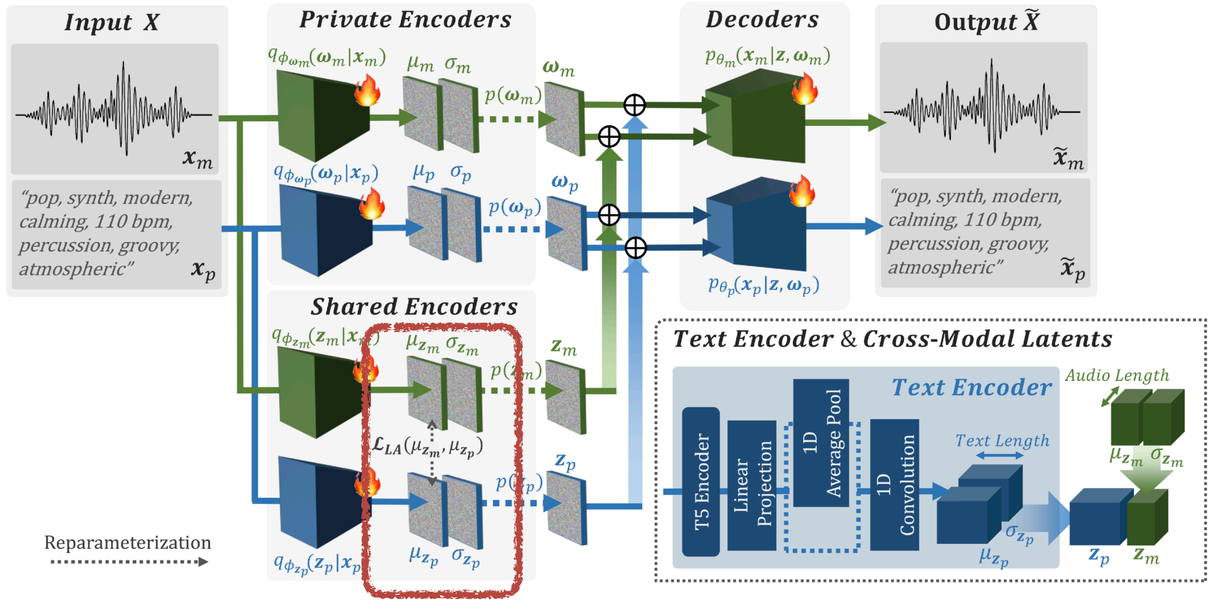
- **MoE-style mVAE Objective for the TAL Space**
  - **Stable Audio Open: audio-only VAE training**
    - $\mathcal{L}_{VAE} = \alpha_{adv}\mathcal{L}_{adv} + \alpha_{mrstft}\mathcal{L}_{mrstft} + \alpha_{kl}\mathcal{L}_{kl}$
      - ** $\mathcal{L}_{adv}$: an adversarial loss
      - ** $\mathcal{L}_{mrstft}$: an audio reconstruction loss
      - ** $\mathcal{L}_{kl}$: a KL divergence term
  - **Our MoE-style mVAE Objective**
    - **Audio Reconstruction:** $\mathcal{L}_{mrstft}$
    - **Text Reconstruction:** $\mathcal{L}_{ce}$



2.2. Text-Aware Latent Space

- **MoE-style mVAE Objective for the TAL Space**
  - **Stable Audio Open: audio-only VAE training**
    - $\mathcal{L}_{VAE} = \alpha_{adv}\mathcal{L}_{adv} + \alpha_{mrstft}\mathcal{L}_{mrstft} + \alpha_{kl}\mathcal{L}_{kl}$
      - ** $\mathcal{L}_{adv}$: an adversarial loss
      - ** $\mathcal{L}_{mrstft}$: an audio reconstruction loss
      - ** $\mathcal{L}_{kl}$: a KL divergence term
  - **Our MoE-style mVAE Objective**
    - **Audio Reconstruction:** $\mathcal{L}_{mrstft}$
    - **Text Reconstruction:** $\mathcal{L}_{ce}$
    - **Latent Alignment (LA) loss** $\mathcal{L}_{LA}$
      - TAL uses the text shared latent z_p

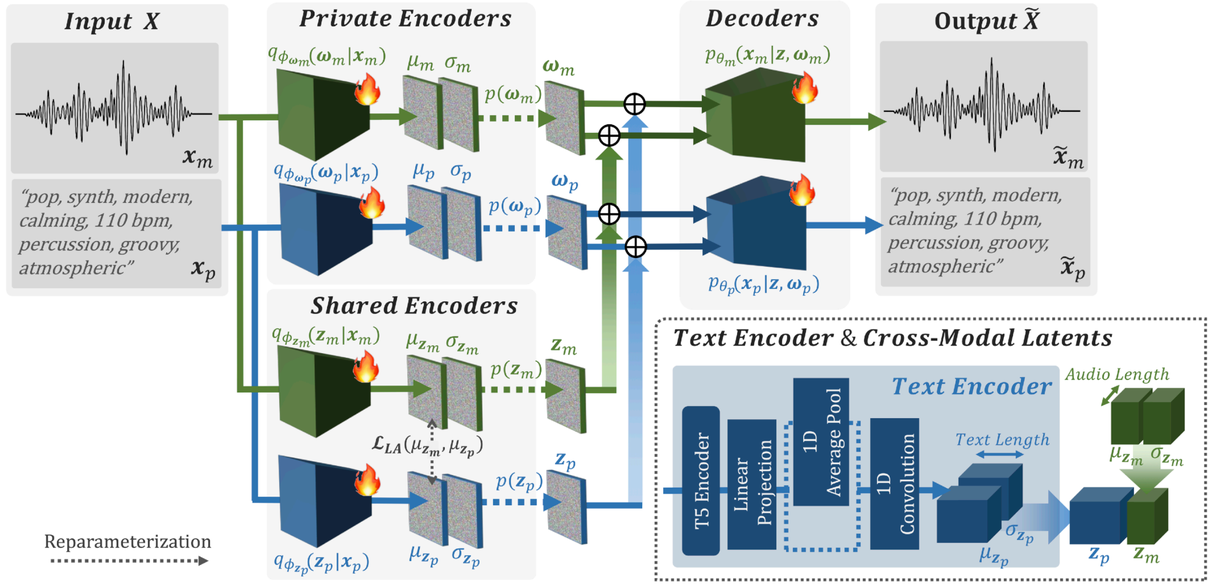
        → z_p must carry "audio-aligned shared information"
      - $L_{LA}(z_m, z_p) = \frac{1}{D}\|\mu_{z_m} - \mu_{z_p}\|_2^2$



2.2. Text-Aware Latent Space

- **MoE-style mVAE Objective for the TAL Space**
  - **Stable Audio Open: audio-only VAE training**
    - $\mathcal{L}_{VAE} = \alpha_{adv}\mathcal{L}_{adv} + \alpha_{mrstft}\mathcal{L}_{mrstft} + \alpha_{kl}\mathcal{L}_{kl}$
      - ** $\mathcal{L}_{adv}$: an adversarial loss
      - ** $\mathcal{L}_{mrstft}$: an audio reconstruction loss
      - ** $\mathcal{L}_{kl}$: a KL divergence term
  - **Our MoE-style mVAE Objective**
    - **Audio Reconstruction:** $\mathcal{L}_{mrstft}$
    - **Text Reconstruction:** $\mathcal{L}_{ce}$
    - **Latent Alignment (LA) loss** $\mathcal{L}_{LA}$
      - TAL uses the text shared latent z_p

        → z_p must carry "audio-aligned shared information"
      - $L_{LA}(z_m, z_p) = \frac{1}{D}\|\mu_{z_m} - \mu_{z_p}\|_2^2$
    - **Final weighted objective**
      - MoE-style reconstruction averaging



$$L_{MoE\text{-}mVAE} = \alpha_{adv}L_{adv} + \alpha_{LA}L_{LA} +$$

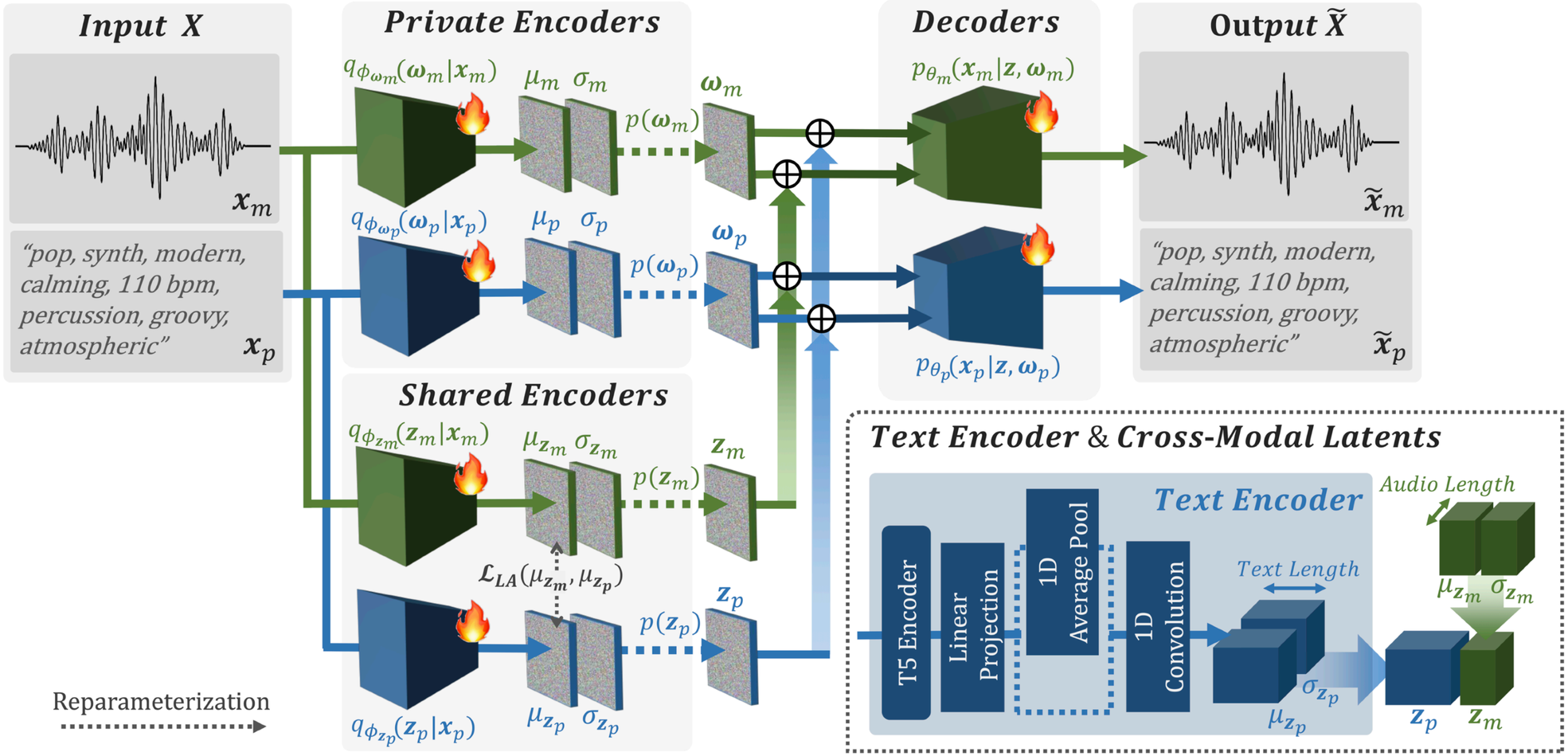
$$\underbrace{\alpha_{mrstft}\left(\frac{1}{2}L_{mrstft}(q_{\phi_{z_m}}, q_{\phi_{\omega_m}}, p_{\theta_m}) + \frac{1}{2}L_{mrstft}(q_{\phi_{z_p}}, q_{\phi_{\omega_m}}, p_{\theta_m})\right)}_{Audio\quad Reconstruction\quad term} +$$

$$\underbrace{\alpha_{ce}\left(\frac{1}{2}L_{ce}(q_{\phi_{z_p}}, q_{\phi_{\omega_t}}, p_{\theta_t}) + \frac{1}{2}L_{ce}(q_{\phi_{z_m}}, q_{\phi_{\omega_t}}, p_{\theta_t})\right)}_{Text\quad Reconstruction\quad term} +$$

$$\underbrace{L_{kl}(q_{\phi_{z_m}}) + L_{kl}(q_{\phi_{z_p}}) + L_{kl}(q_{\phi_{\omega_m}}) + L_{kl}(q_{\phi_{\omega_t}})}_{Regularization\quad term}$$

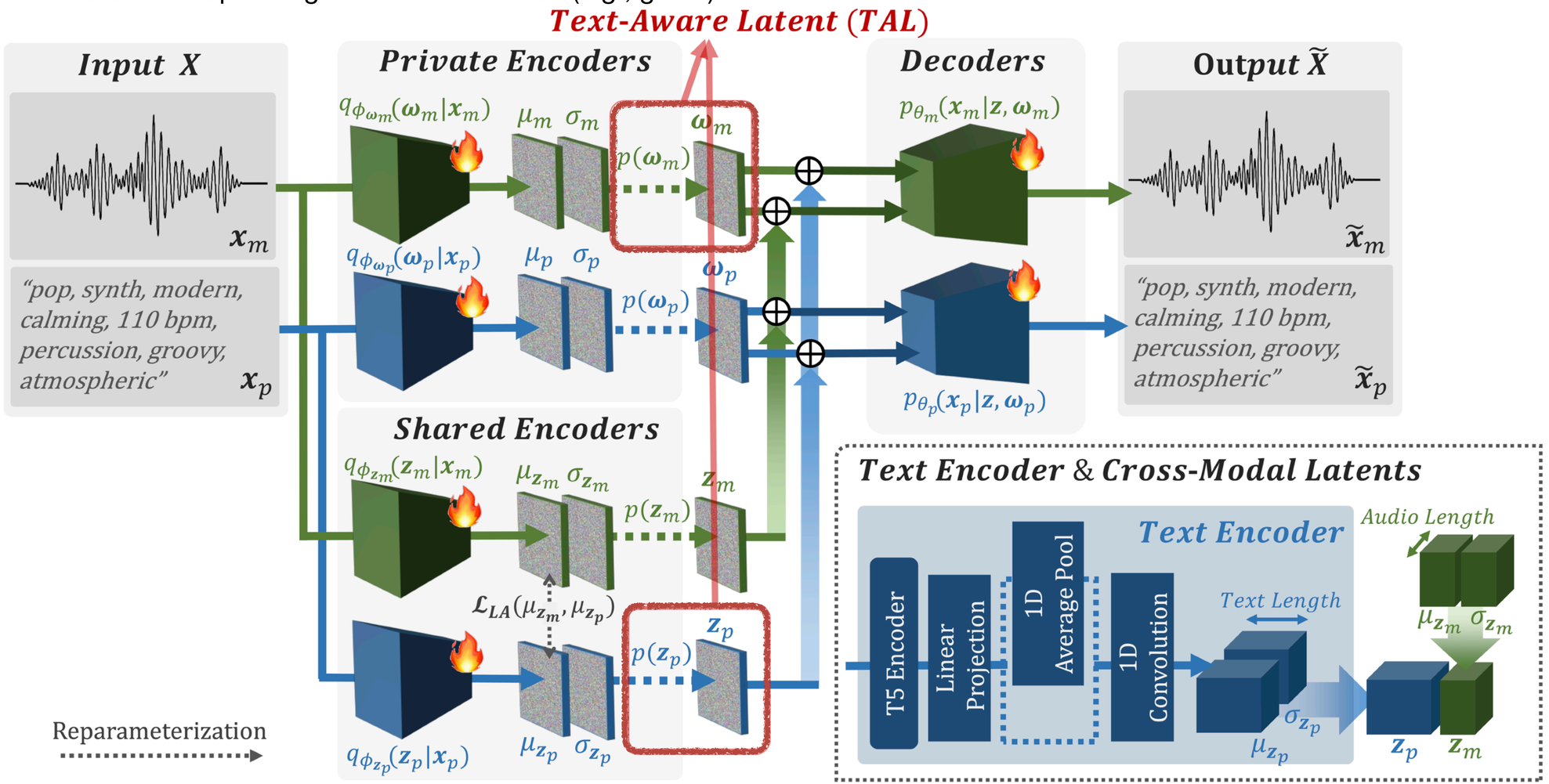
2.2. Text-Aware Latent Space

- **TAL Definition :** $[w_m; z_p]$
  - w_m : private audio latent → preserves audio reconstruction quality
  - z_p : shared text latent → captures global text semantics (e.g., genre)



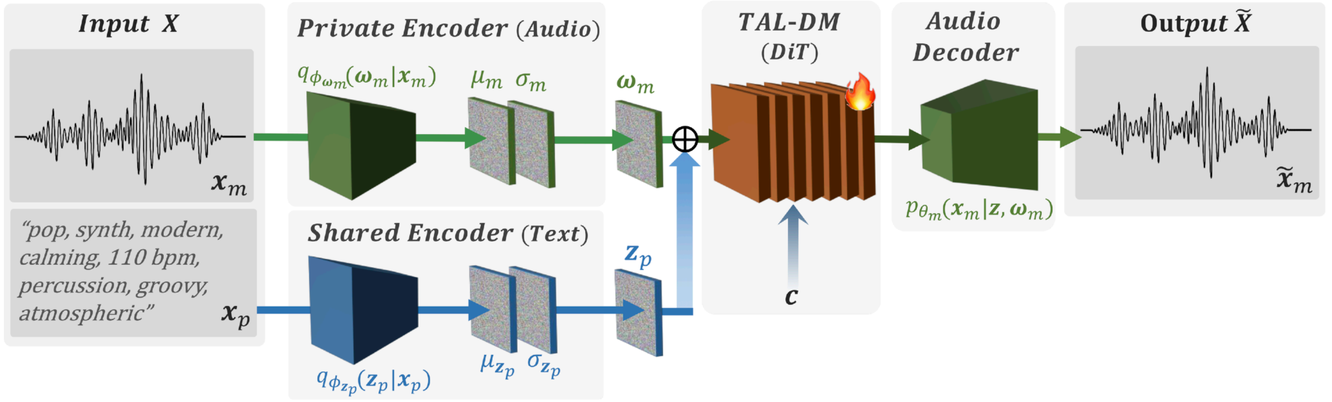
2.2. Text-Aware Latent Space

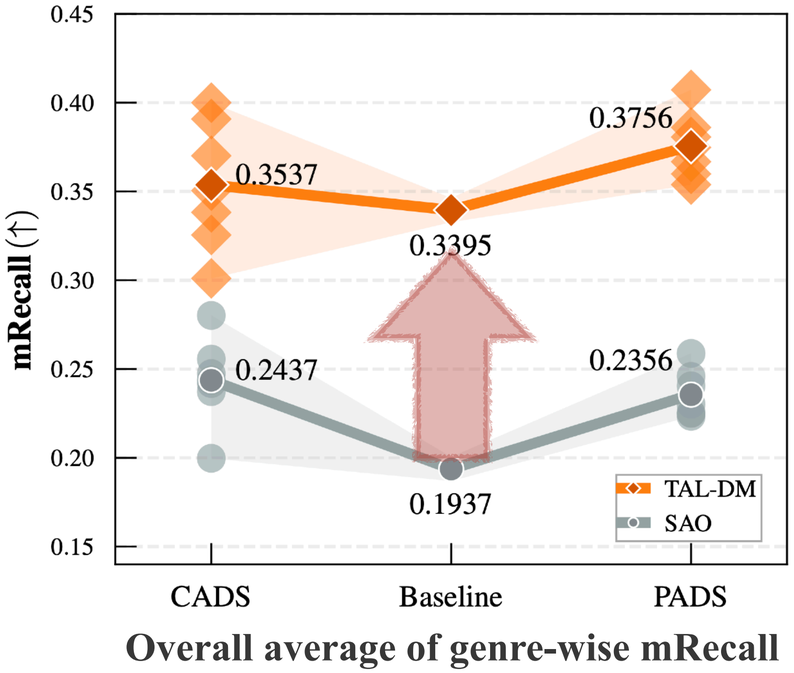
- **TAL Definition :** $[w_m; z_p]$
  - w_m : private audio latent → preserves audio reconstruction quality
  - z_p : shared text latent → captures global text semantics (e.g., genre)



2.3. Unified Pipeline

- **After Training a Diffusion Model in the TAL Space**
  - **Sampling with TAL-DM**
    - within text-conditioned latent neighborhoods
    - strengthens global consistency
    - mitigates genre mismatch

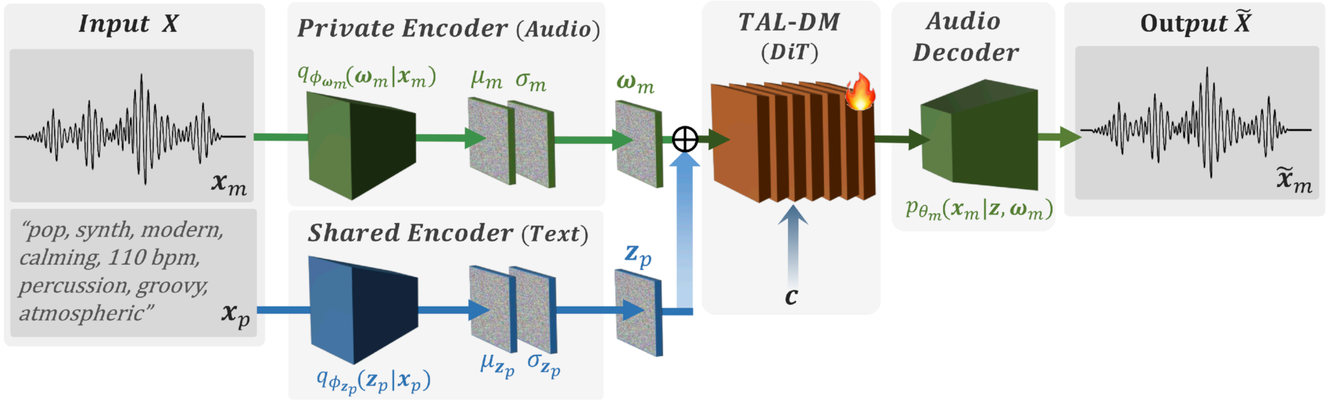


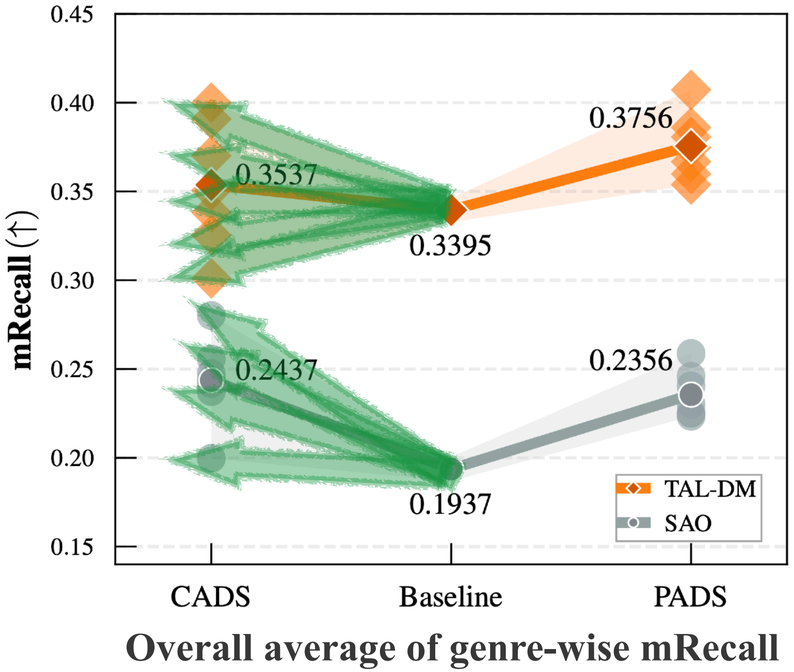


2.3. Unified Pipeline

- **After Training a Diffusion Model in the TAL Space**
  - **Sampling with TAL-DM**
    - within text-conditioned latent neighborhoods
    - strengthens global consistency
    - mitigates genre mismatch

- **Complementary Effects in the Unified Pipeline**
  - **CADS**
    - may uncover new reference-manifold regions
    - but gains are erratic across perturbation settings

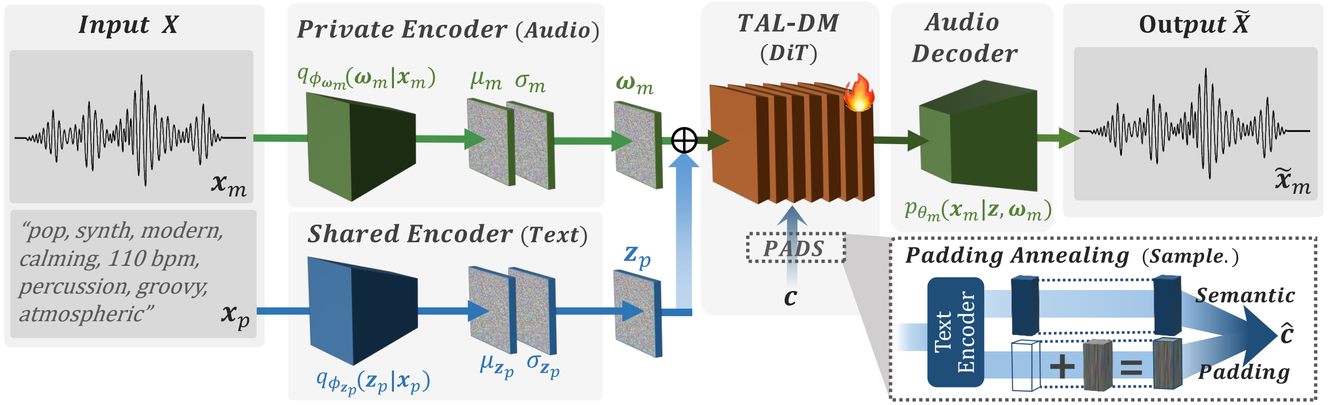


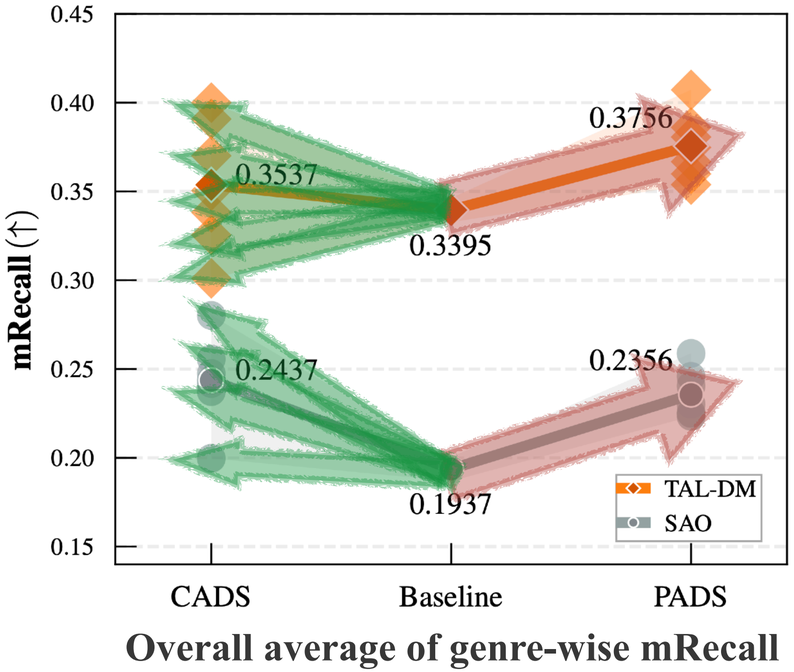


2.3. Unified Pipeline

- **After Training a Diffusion Model in the TAL Space**
  - **Sampling with TAL-DM**
    - within text-conditioned latent neighborhoods
    - strengthens global consistency
    - mitigates genre mismatch

- **Complementary Effects in the Unified Pipeline**
  - **CADS**
    - may uncover new reference-manifold regions
    - but gains are erratic across perturbation settings
  - **PADS in TAL space**
    - suppresses direct corruption of semantic embeddings
    - stably expands genre-consistent diversity





3. Experiments

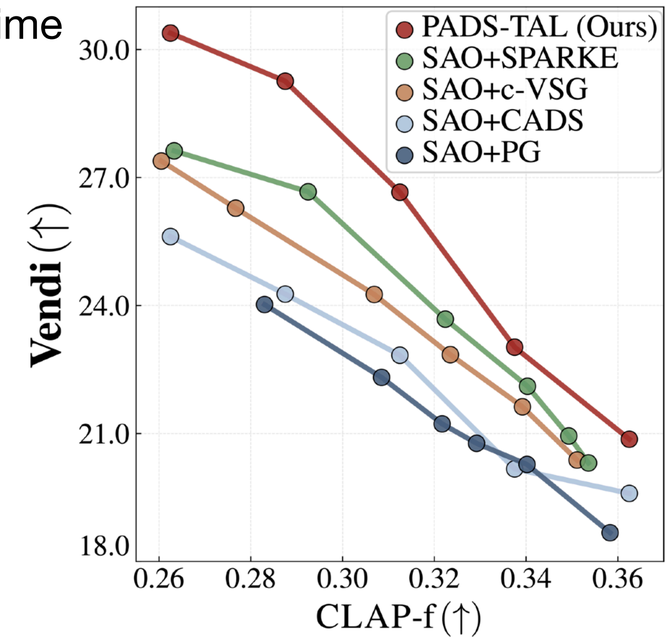
- **Main Results**
  - **Ablation: TAL and PADS**
    - Matched CLAP-f (≈0.32)
    - PADS-TAL achieves best
    - Key gain over “Audio + CADS”
      - +15.4% overall diversity
      - +71.6% within-genre diversity

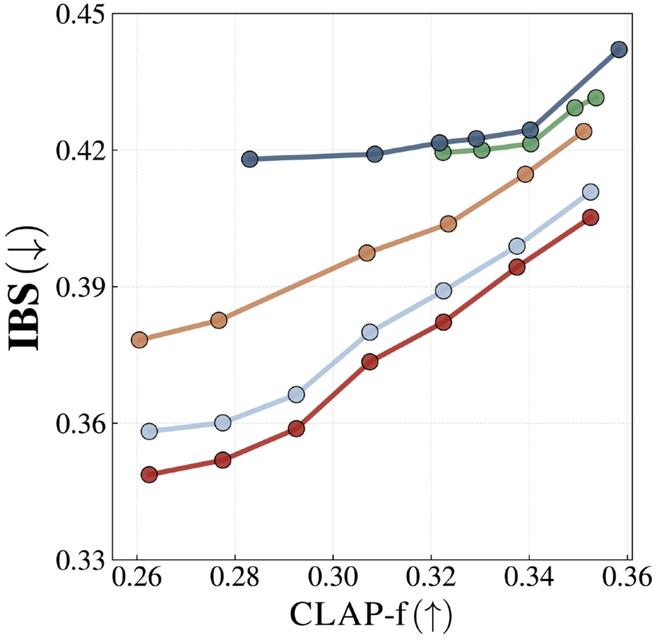
Architecture		Evaluation Dataset			
Latent Space	Pert.	Melbench mRecall↑	SongDescriber Vendi ↑	SongDescriber KL_{passt} ↓	SongDescriber FD_{openl3} ↓
Audio	**CADS**	0.237	22.591	0.669	147.730
Audio	**PADS**	0.212	24.367	0.652	139.838
TAL	**CADS**	0.391	23.873	0.669	145.393
TAL	**PADS**	**0.407**	**26.075**	**0.626**	**135.542**

3. Experiments

- **Main Results**
  - **Ablation: TAL and PADS**
    - Matched CLAP-f (≈0.32)
    - PADS-TAL achieves best
    - Key gain over “Audio + CADS”
      - +15.4% overall diversity
      - +71.6% within-genre diversity
  - **Comparison with Diversity Methods**
    - PADS-TAL : higher diversity in usable alignment regime

Architecture		Evaluation Dataset			
Latent Space	Pert.	Melbench mRecall↑	SongDescriber Vendi ↑	SongDescriber KL_{passt} ↓	SongDescriber FD_{openl3} ↓
Audio	**CADS**	0.237	22.591	0.669	147.730
Audio	**PADS**	0.212	24.367	0.652	139.838
TAL	**CADS**	0.391	23.873	0.669	145.393
TAL	**PADS**	**0.407**	**26.075**	**0.626**	**135.542**





3. Experiments

- **Main Results**
  - **Ablation: TAL and PADS**
    - Matched CLAP-f (≈0.32)
    - PADS-TAL achieves best
    - Key gain over “Audio + CADS”
      - +15.4% overall diversity
      - +71.6% within-genre diversity
  - **Comparison with Diversity Methods**
    - PADS-TAL : higher diversity in usable alignment regime
  - **Human Evaluation**
    - PADS-TAL: highest diversity with preserved quality/alignment

Architecture		Evaluation Dataset			
Latent Space	Pert.	Melbench mRecall↑	SongDescriber Vendi ↑	SongDescriber KL_{passt} ↓	SongDescriber FD_{openl3} ↓
Audio	**CADS**	0.237	22.591	0.669	147.730
Audio	**PADS**	0.212	24.367	0.652	139.838
TAL	**CADS**	0.391	23.873	0.669	145.393
TAL	**PADS**	**0.407**	**26.075**	**0.626**	**135.542**

(a)

	Quality	Diversity	Alignment
SAO+CADS	3.18	3.47	3.09
SAO+PADS	**3.89**	**3.49**	**3.79**

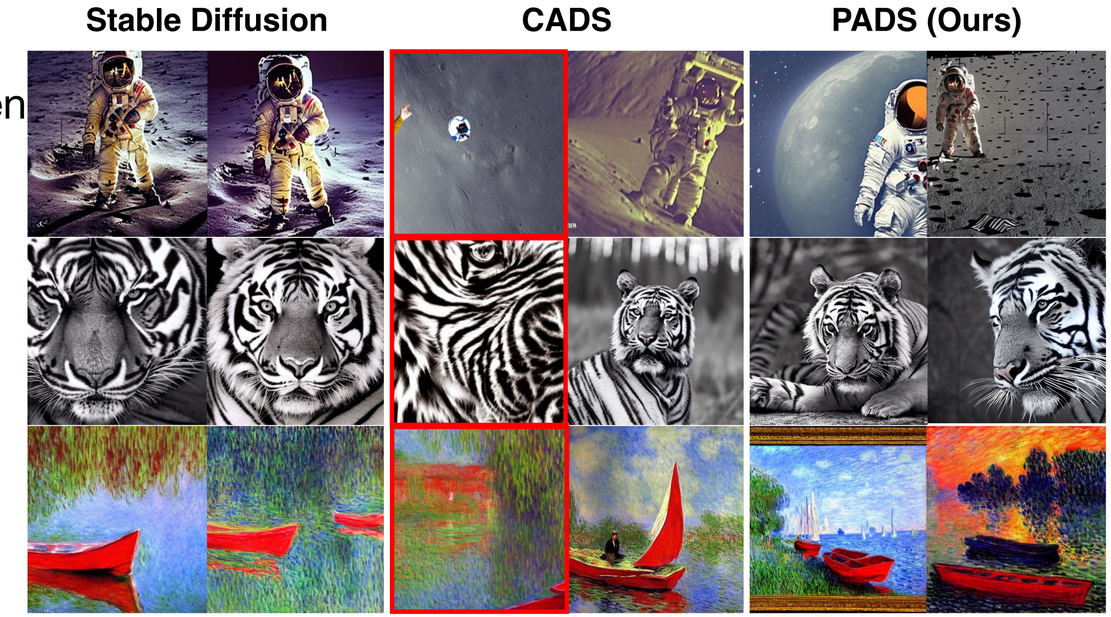
(b)

	Quality	Diversity	Alignment
SAO	**3.81**	3.41	**3.83**
SAO+CADS	3.36	3.56	2.87
PADS-TAL	3.79	**3.71**	3.78

3. Experiments

- **Main Results**
  - **Ablation: TAL and PADS**
    - Matched CLAP-f (≈0.32)
    - PADS-TAL achieves best
    - Key gain over “Audio + CADS”
      - +15.4% overall diversity
      - +71.6% within-genre diversity
  - **Comparison with Diversity Methods**
    - PADS-TAL : higher diversity in usable alignment regime
  - **Human Evaluation**
    - PADS-TAL: highest diversity with preserved quality/alignmen
- **Generalization of PADS**
  - **Across T2M frameworks**
    - CADS: hurts alignment or fails to improve diversity
    - PADS: stable diversity gain
  - **Beyond T2M : T2I examples**
    - CADS: occasional misalignment
    - PADS: diversity with maintained alignment

Model	Pert.	Vendi ↑	CLAP-f↑	KL_{passt}↓
ACE-Step	-	16.6154	0.2425	0.7456
	CADS	**23.4070** △ +6.7916	0.1981 ▽ –0.0444	0.9240 △ +0.1784
	PADS	**23.3399** △ +6.7245	0.2318 ▽ –0.0107	0.7918 △ +0.0462
MelodyFlow	-	18.2190	0.2770	0.6873
	PADS	**21.5843** △ +3.3653	0.2557 ▽ –0.0213	0.7671 △ +0.0798



Q & A