深度伪造检测（Deepfake Detection）

—— 研究进展汇报

汇报人：邱俊航

DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion

Ke Sun[1], **Shen Chen**[2], **Taiping Yao**[2], **Hong Liu**[3]*,
Xiaoshuai Sun[1], **Shouhong Ding**[2], **Rongrong Ji**[1]
[1] Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China.
[2] Youtu Lab, Tencent, P.R. China.
[3] Osaka University, Japan.

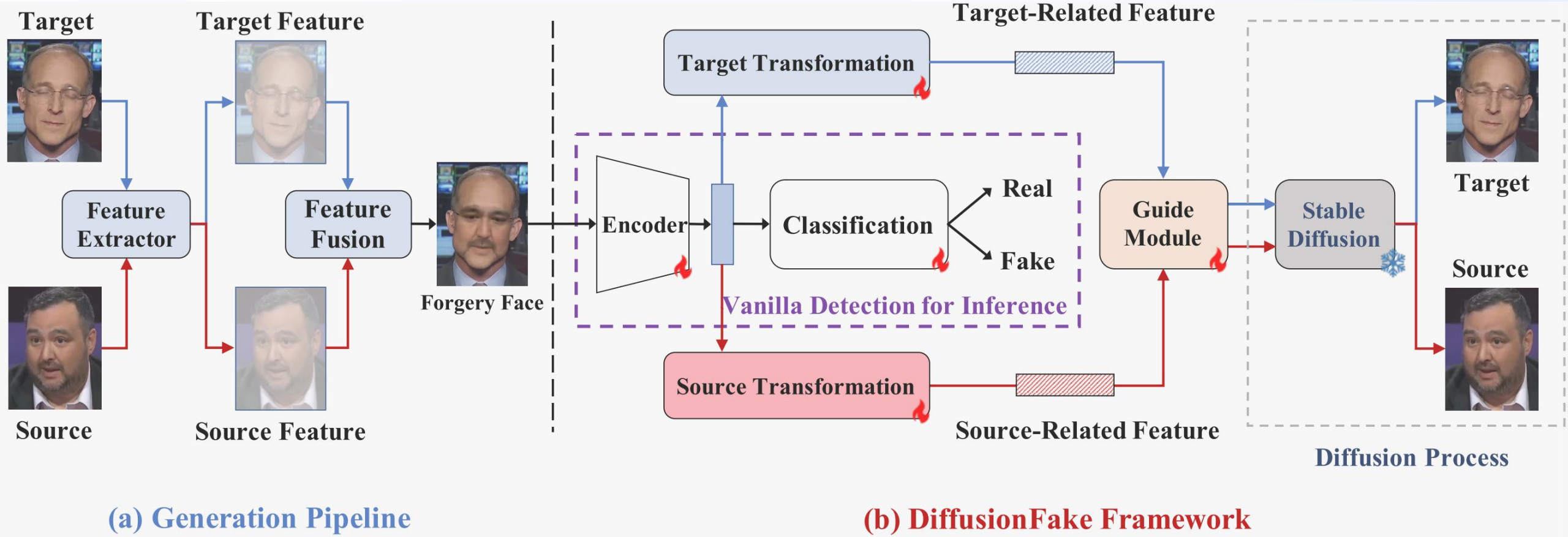
NeurIPS 2024

Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake

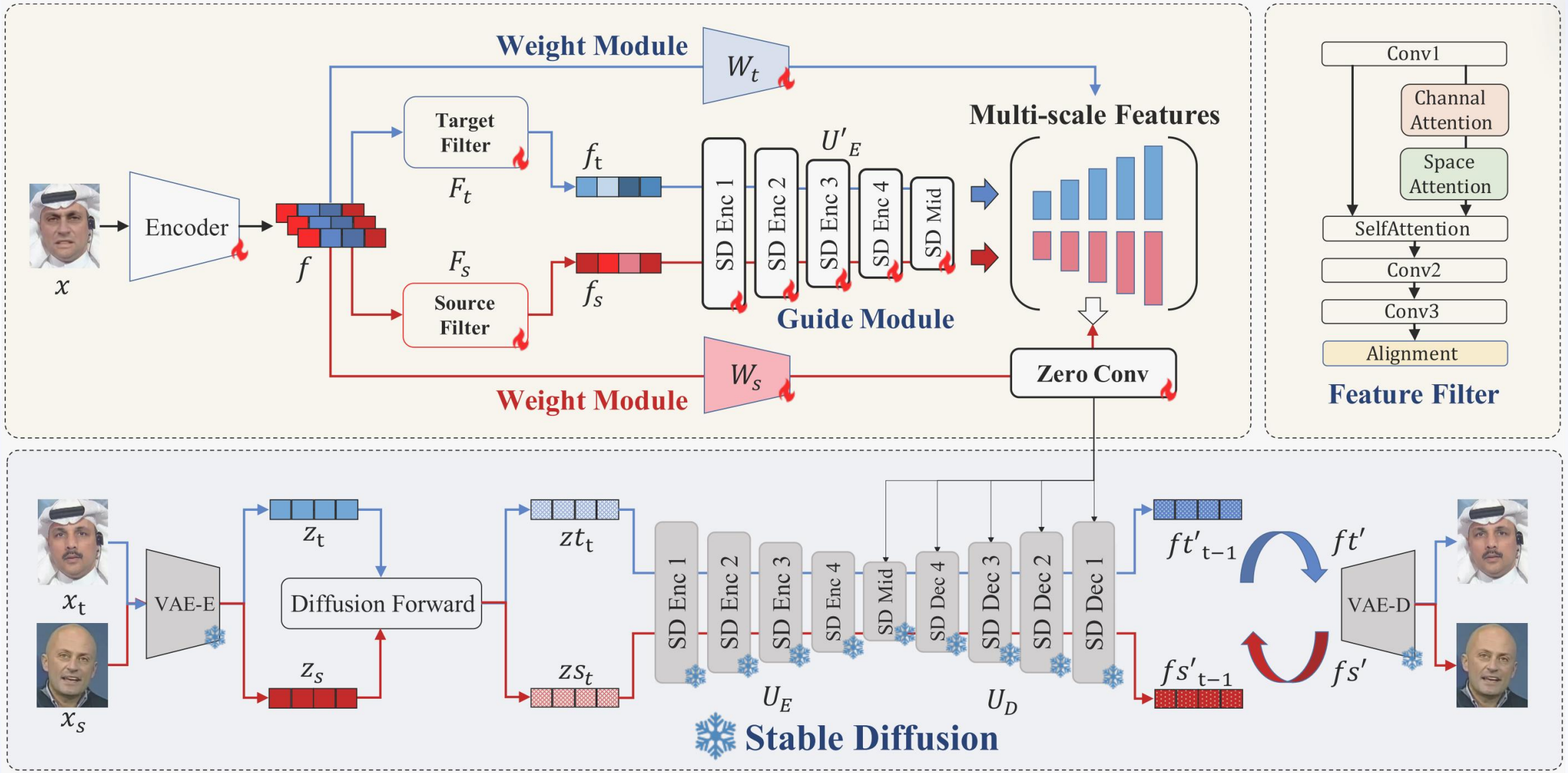
文章提出：Deepfake图像本质上融合了来自源和目标面部的信息，而真正的图像始终保持一致的身份。

通过使用一个预先训练的强大模型（Stable-Diffusion）来指导 BackBone 的模型提取图像特征，在这篇文章中，Feature-moudle、Weight-module 、 Guide-moudle 和 Stable-Diffusion 共同提高了编码器和分类器的能力。



Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- 即插即用的训练框架



Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- Feature Filter moudle

特征过滤模型负责从 Encoder 编码后的特征中提取源相关特征和目标相关特征，然后送入指导模型中。

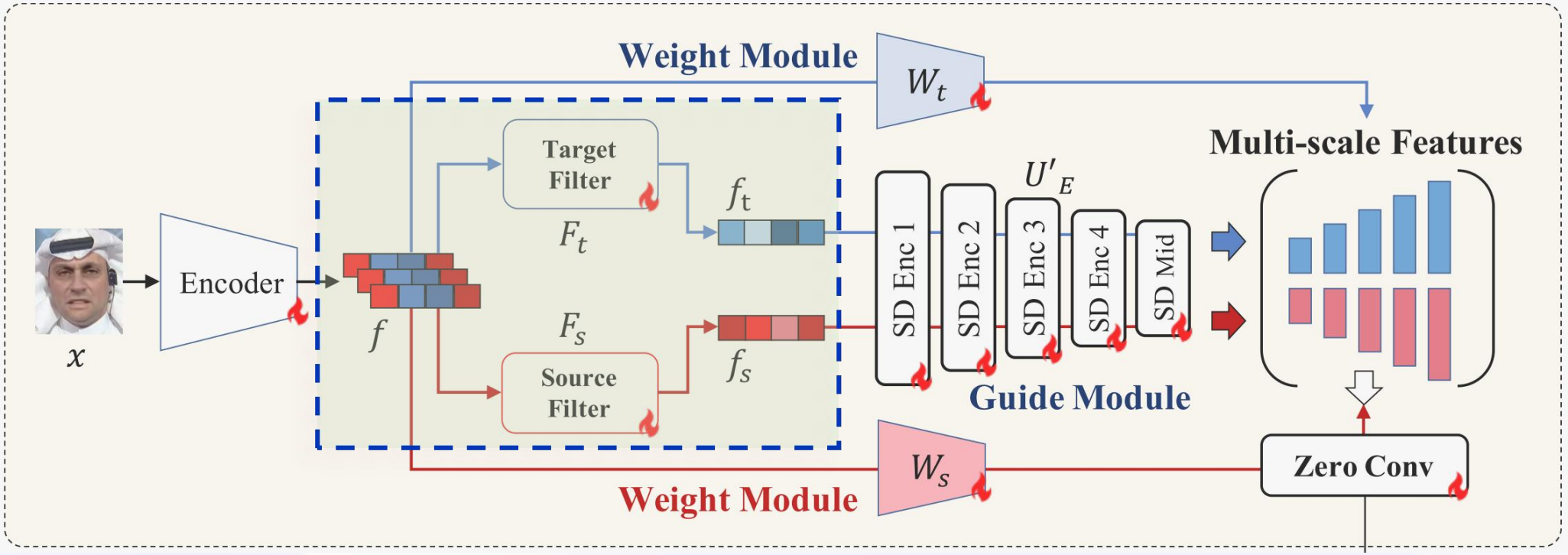
特征过滤模型（Feature Filter Module）由以下模块构成：

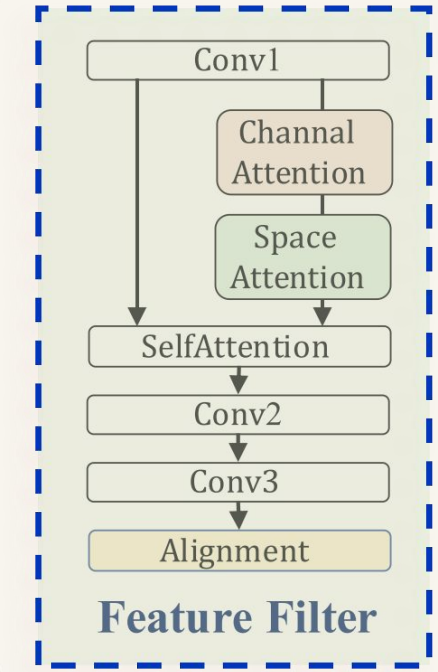
1）通道式注意力（Channel Attention）

2）空间式注意力（Space Attention）

3）多头自注意力（Multi-Self Attention）

4）上采样和池化（Upsample & Pooling）

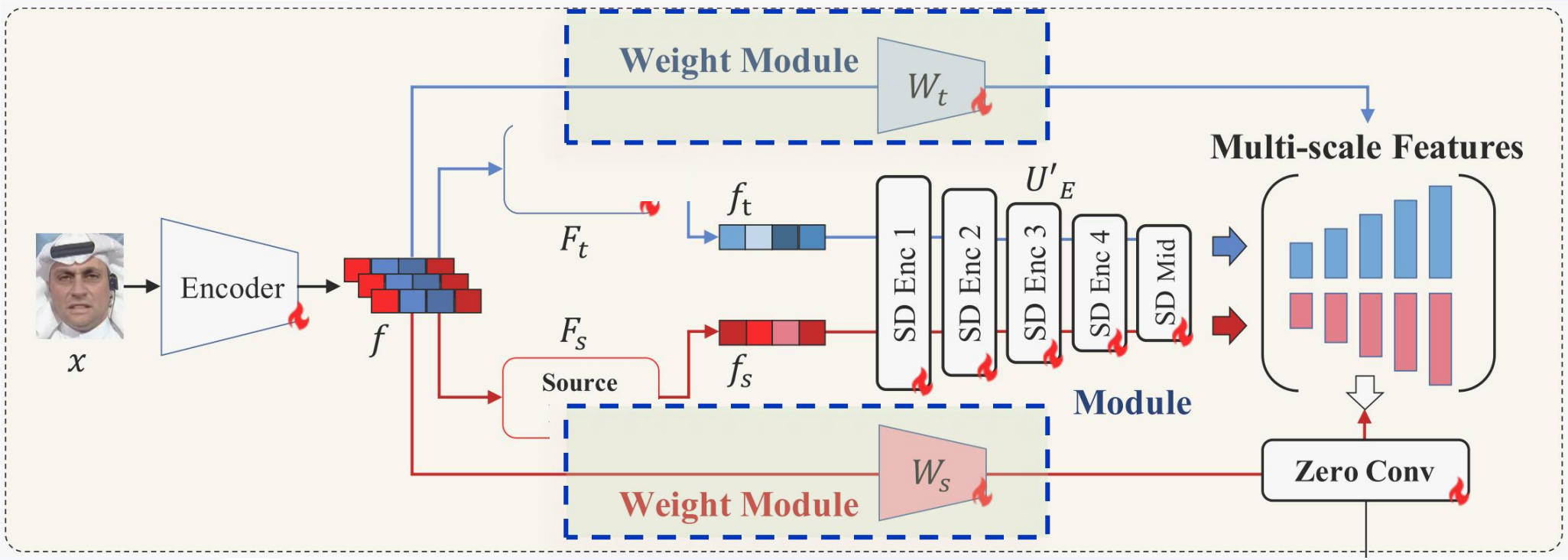




Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- Weight moudle

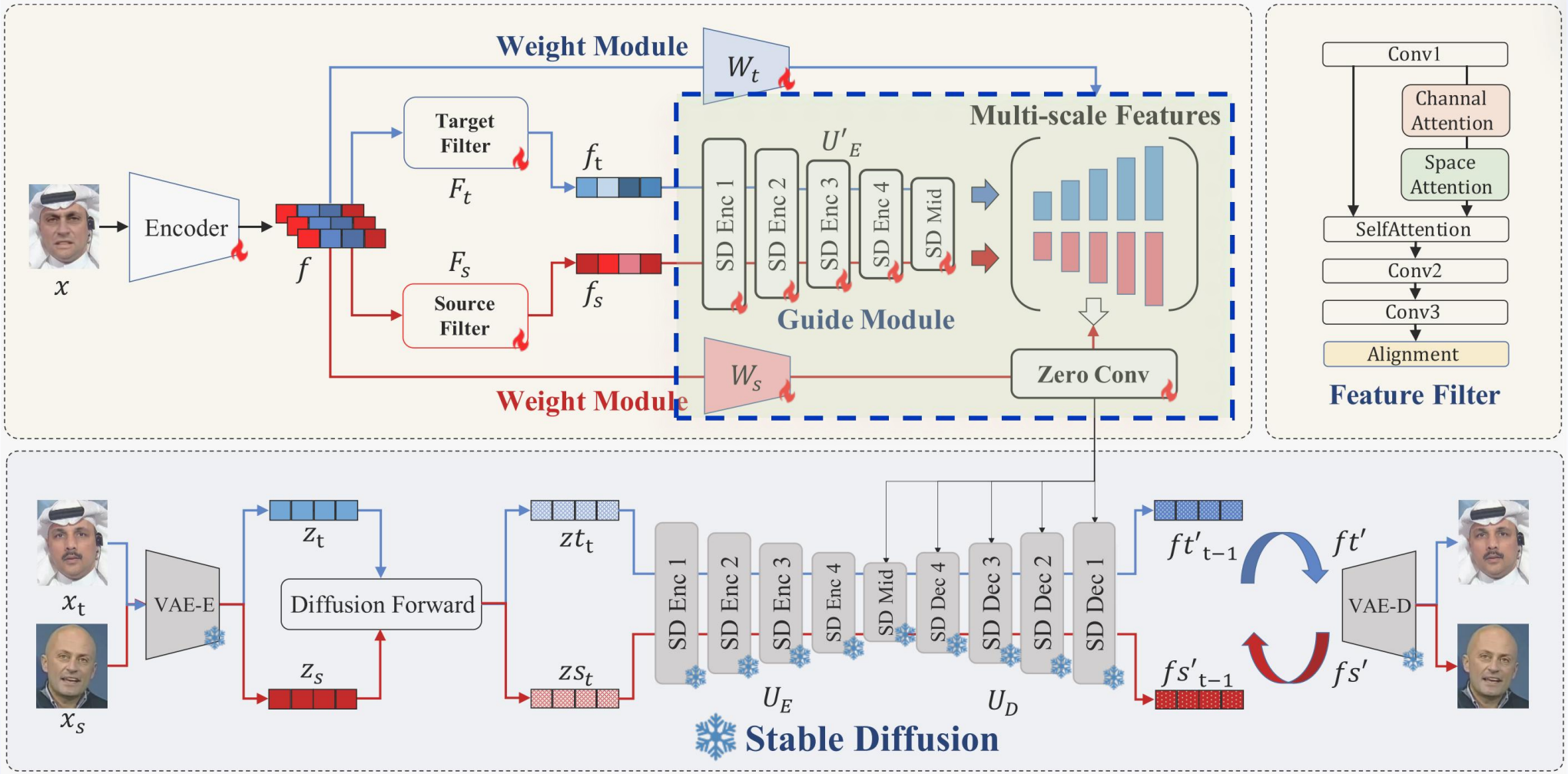
对 Encoder 提取出的 Target 特征和 Source 特征的占比进行动态调整，确定二者融入最终多尺度特征（Multi-scale Features）的权重。



Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- Guide moudle

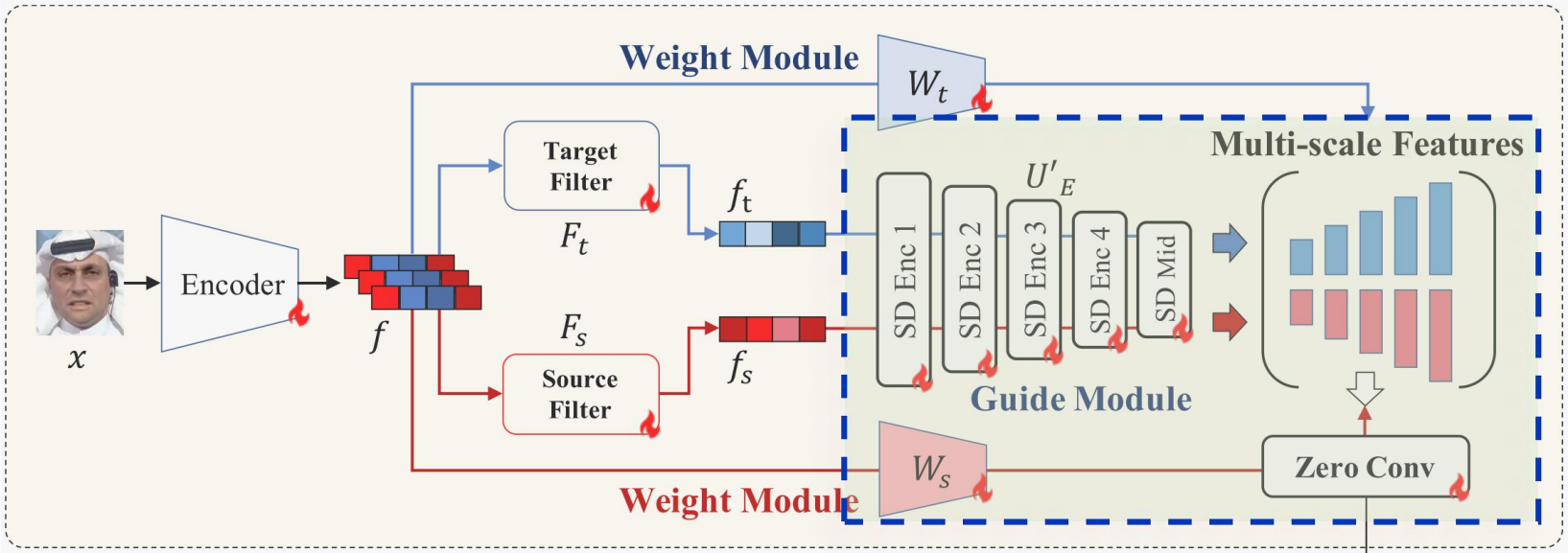
指导模块的架构与 Stable-Diffusion 的编码模块相同，指导模块把特征过滤模块得到的两组特征进行再编码然后输入到零卷积层，随后将两个特征与 weight-module 给到的权重乘积得到最终的输出特征。

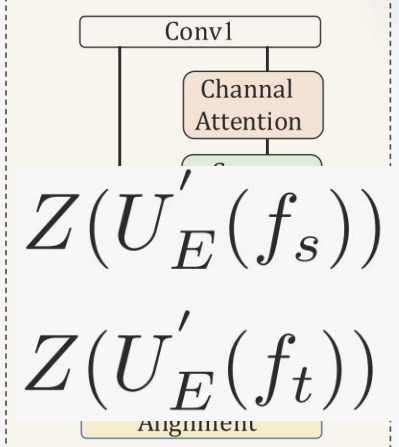


Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- Guide moudle

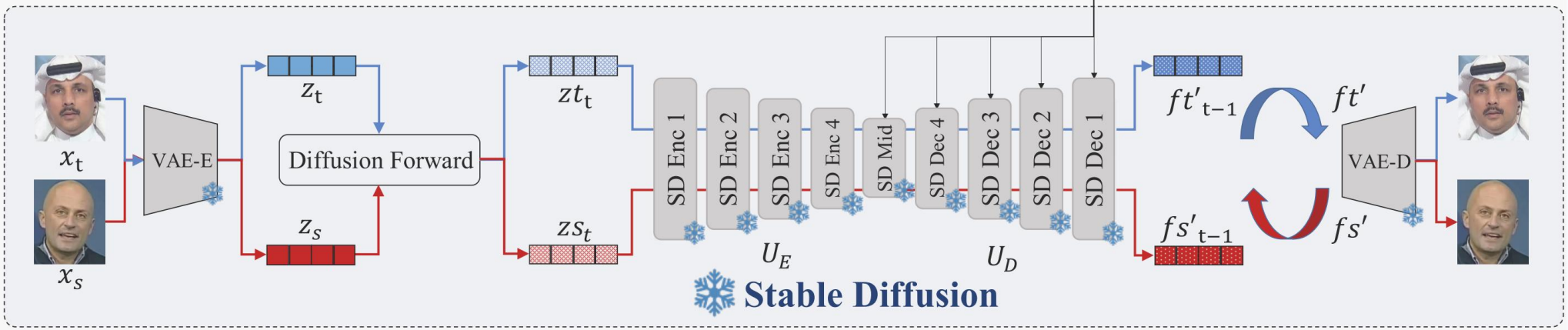
指导模块的架构与 Stable-Diffusion 的编码模块相同，指导模块把特征过滤模块得到的两组特征进行再编码然后输入到零卷积层，随后将两个特征与 weight-module 给到的权重乘积得到最终的输出特征。





$$Z(U'_E(f_s)) \times W_s(f)$$

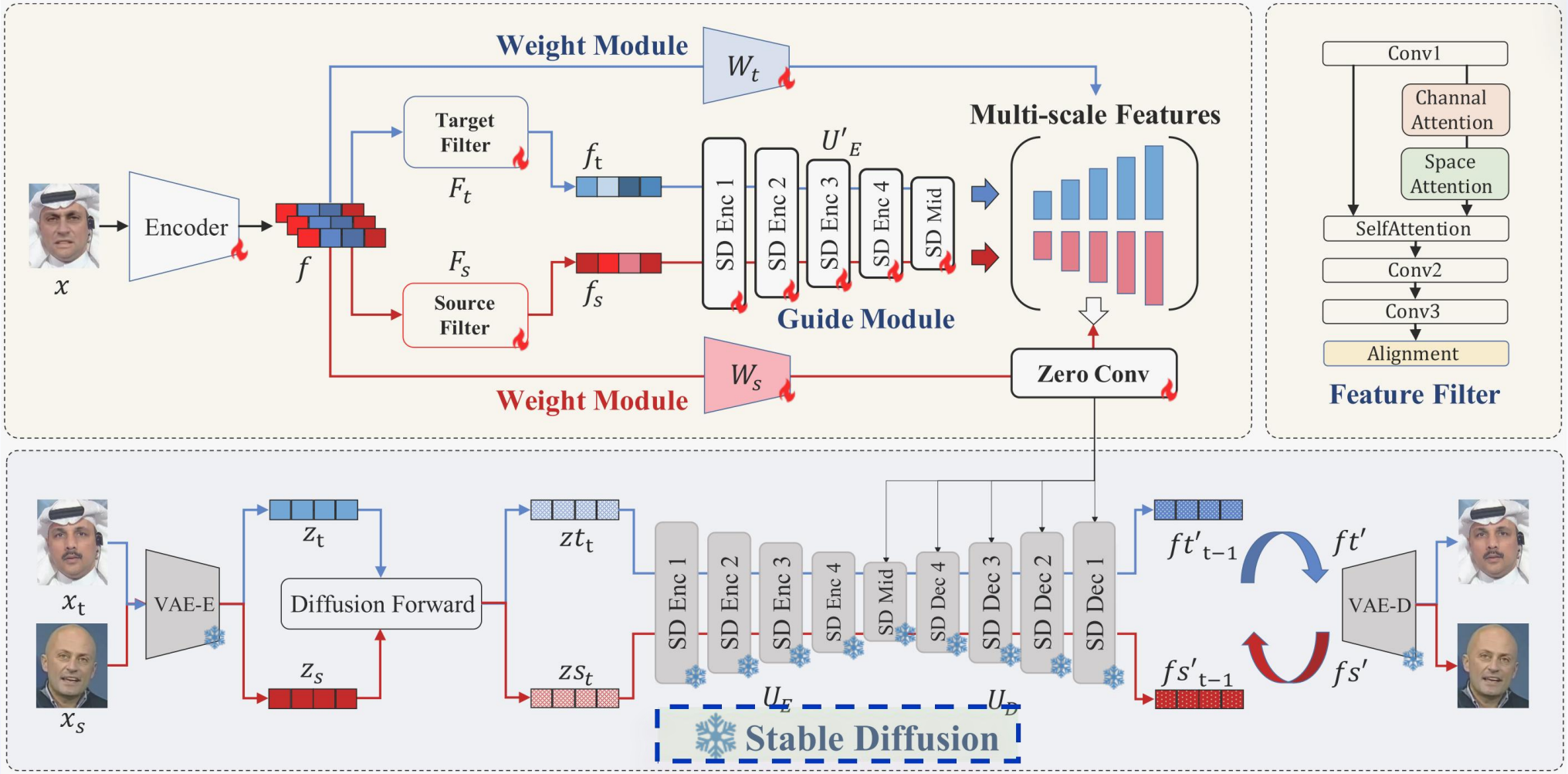
$$Z(U'_E(f_t)) \times W_t(f)$$



Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- Stable Diffusion

VAE 模块对两个图像（源图像和目标图像）提取后的特征输入给 Stable-Diffusion 的编码器，再输入解码器得到潜在特征。Stable-Diffusion 与 Guide-module 提取的特征合并作为后续生成源图像和目标图像的真正 Laten-information。

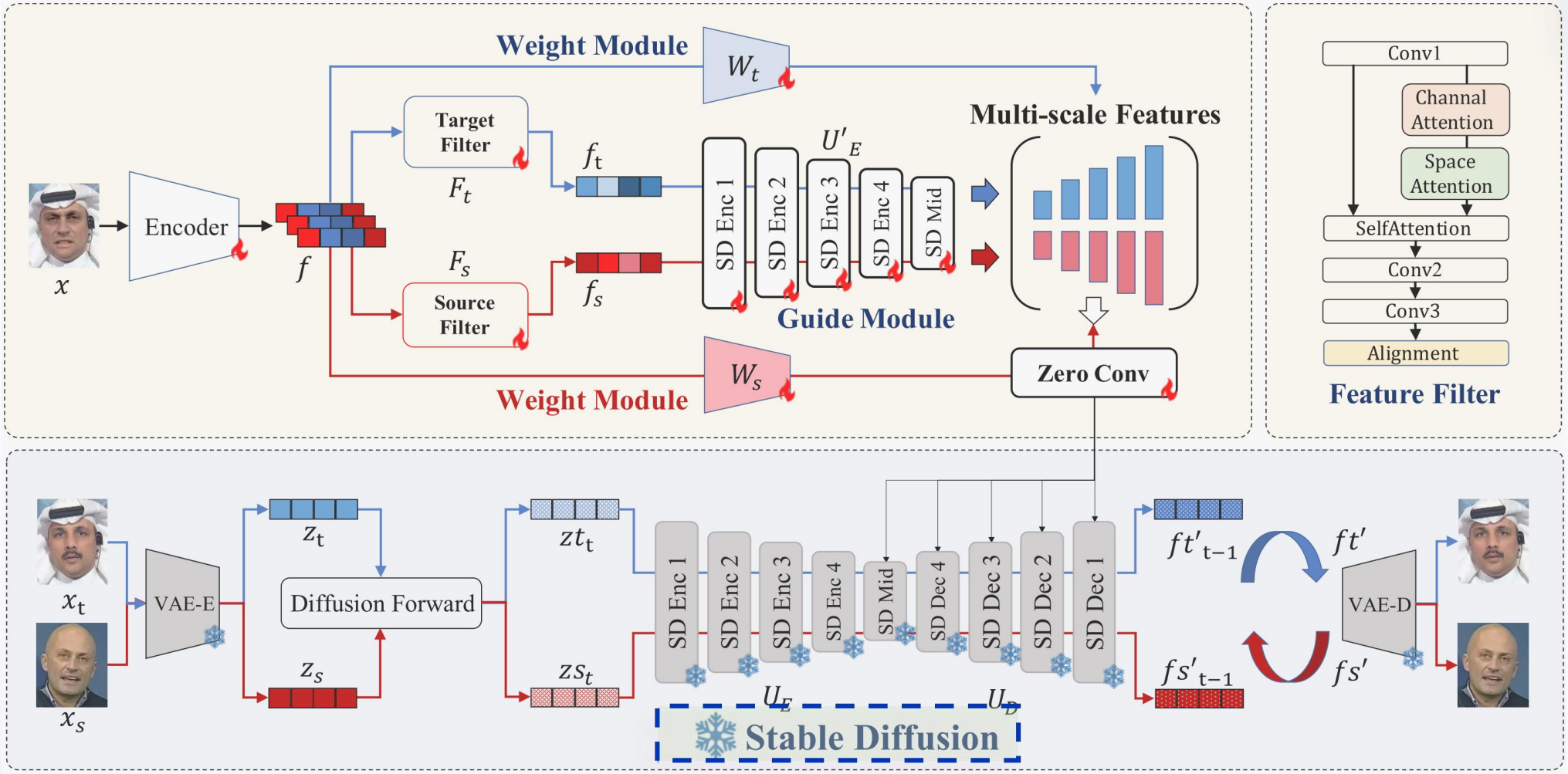


Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- Stable Diffusion

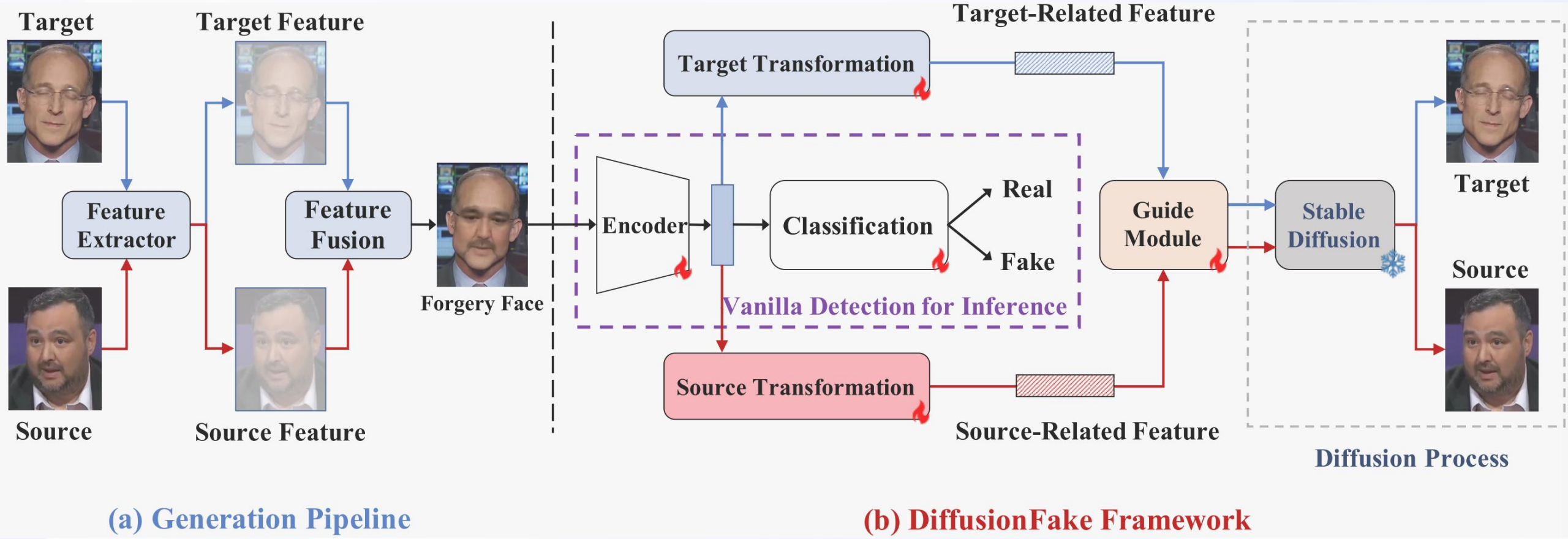
$$fs^{'} = U_D(U_E(z_s)) + Z(U_E^{'}(f_s)) \times W_s(f)$$

$$ft^{'} = U_D(U_E(z_t)) + Z(U_E^{'}(f_t)) \times W_t(f)$$



Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- Stable Diffusion



FF++ 数据集划分规则：

验证和测试精度是在每个视频100张图像上计算的，训练是在每个视频270张图像上评估的，伪图像的数量大约是真实图像数量的四倍。

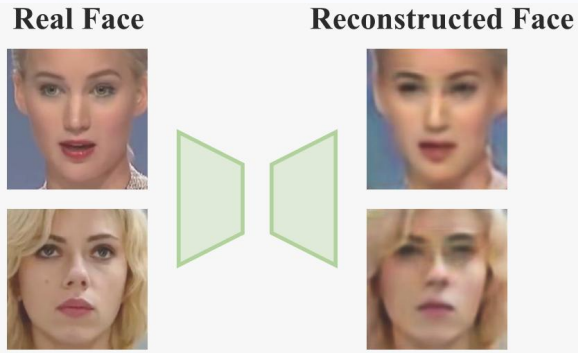
Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

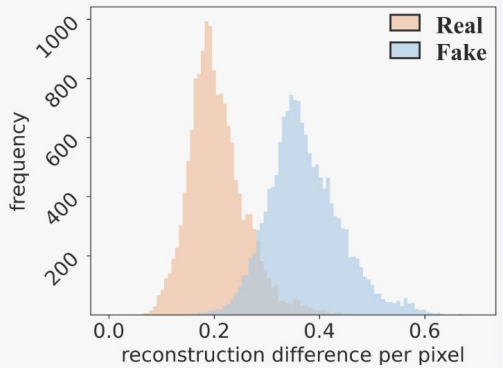
DiffusionFake -- 成效

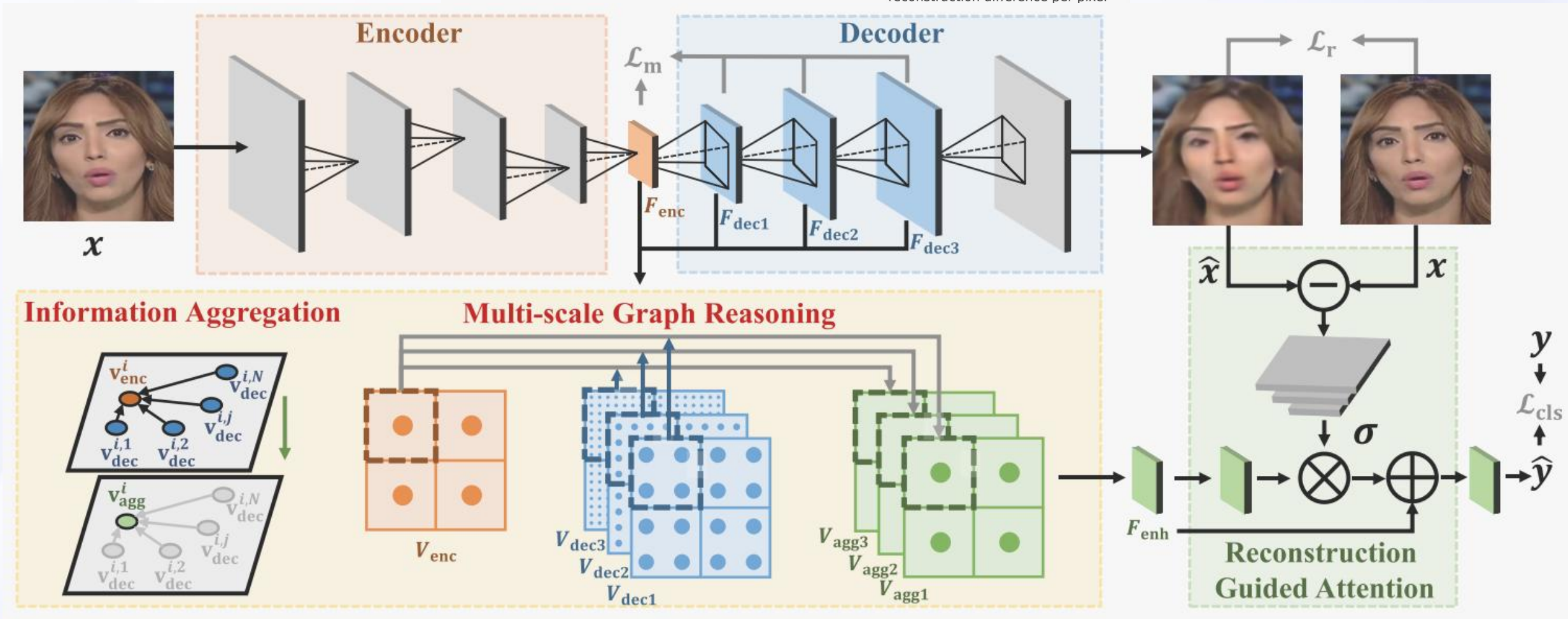
Method	Celeb-DF		Wild Deepfake		DFDC-P		DFD		DiffSwap		Average	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Xception [7]	65.27	38.77	66.17	40.14	69.80	35.41	87.86	21.04	74.25	32.04	72.67	33.48
Face X-ray [42]	74.20	-	-	-	70.00	-	85.60	-	-	-	-	-
F3-Net* [29]	71.21	34.03	67.71	40.17	72.88	33.38	86.10	26.17	76.89	30.83	74.96	32.92
MAT* [50]	70.65	35.83	70.15	36.53	67.34	38.31	87.58	21.73	79.93	27.77	75.13	32.03
GFF* [28]	75.31	32.48	66.51	41.52	71.58	34.77	85.51	25.64	78.38	28.15	75.46	32.51
LTW [40]	77.14	29.34	67.12	39.22	74.58	33.81	88.56	20.57	77.95	29.01	77.07	30.39
LRL [4]	78.26	29.67	68.76	37.50	76.53	32.41	89.24	20.32	-	-	-	-
DCL [41]	82.30	26.53	71.14	36.17	76.71	31.97	91.66	16.63	80.21	27.37	80.40	27.73
PCL+I2G [51]	81.80	-	-	-	-	-	-	-	-	-	-	-
SBI* [35]	80.76	26.97	68.22	38.11	76.53	30.22	88.13	17.25	75.20	31.49	77.77	28.81
UIA-ViT [53]	82.41	-	-	-	75.80	-	**94.68**	-	-	-	-	-
RECCE* [2]	70.50	35.34	67.93	39.82	75.88	32.41	89.91	19.95	77.59	29.38	76.36	31.38
UCF [48]	75.27	-	-	-	75.94	-	80.74	-	-	-	-	-
CADDM* [9]	77.56	30.63	72.56	33.63	72.45	33.56	82.90	25.20	75.58	31.01	76.21	30.81
EN-b4* [42]	73.51	34.17	70.04	37.03	70.51	33.98	87.57	21.31	77.38	29.44	75.80	31.19
VIT-B* [42]	74.64	33.07	75.46	31.53	74.24	34.29	84.38	24.15	78.50	28.14	77.44	30.24
En-b4+Ours	**83.17**	**24.59**	75.17	33.25	77.35	30.17	91.71	**16.27**	82.02	25.55	81.88	25.97
VIT-B+Ours	80.46	27.51	**80.14**	**29.62**	**80.95**	**27.66**	90.36	19.73	**86.98**	**21.32**	**83.78**	**25.17**

Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

RECCE







Cao J, Ma C, Yao T, et al. End-to-end reconstruction-classification learning for face forgery detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4113-4122.

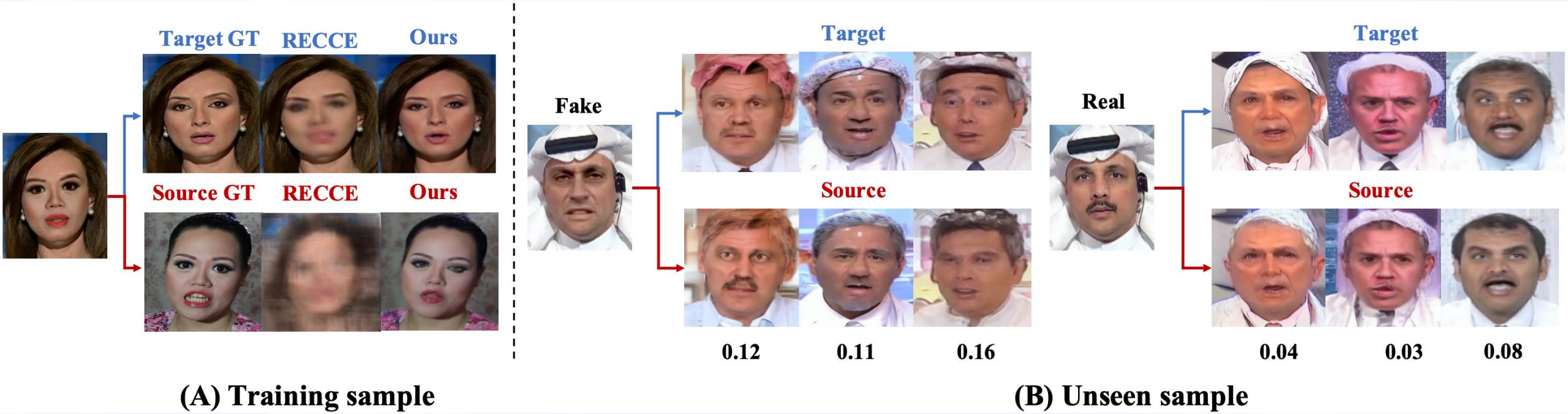
DiffusionFake -- 成效

1、对于训练过的图像，DiffusionFake能够有效的区分出目标图像和源图像；

2、对于未见过的图像，DiffusionFake重建出的图像符合预设条件：

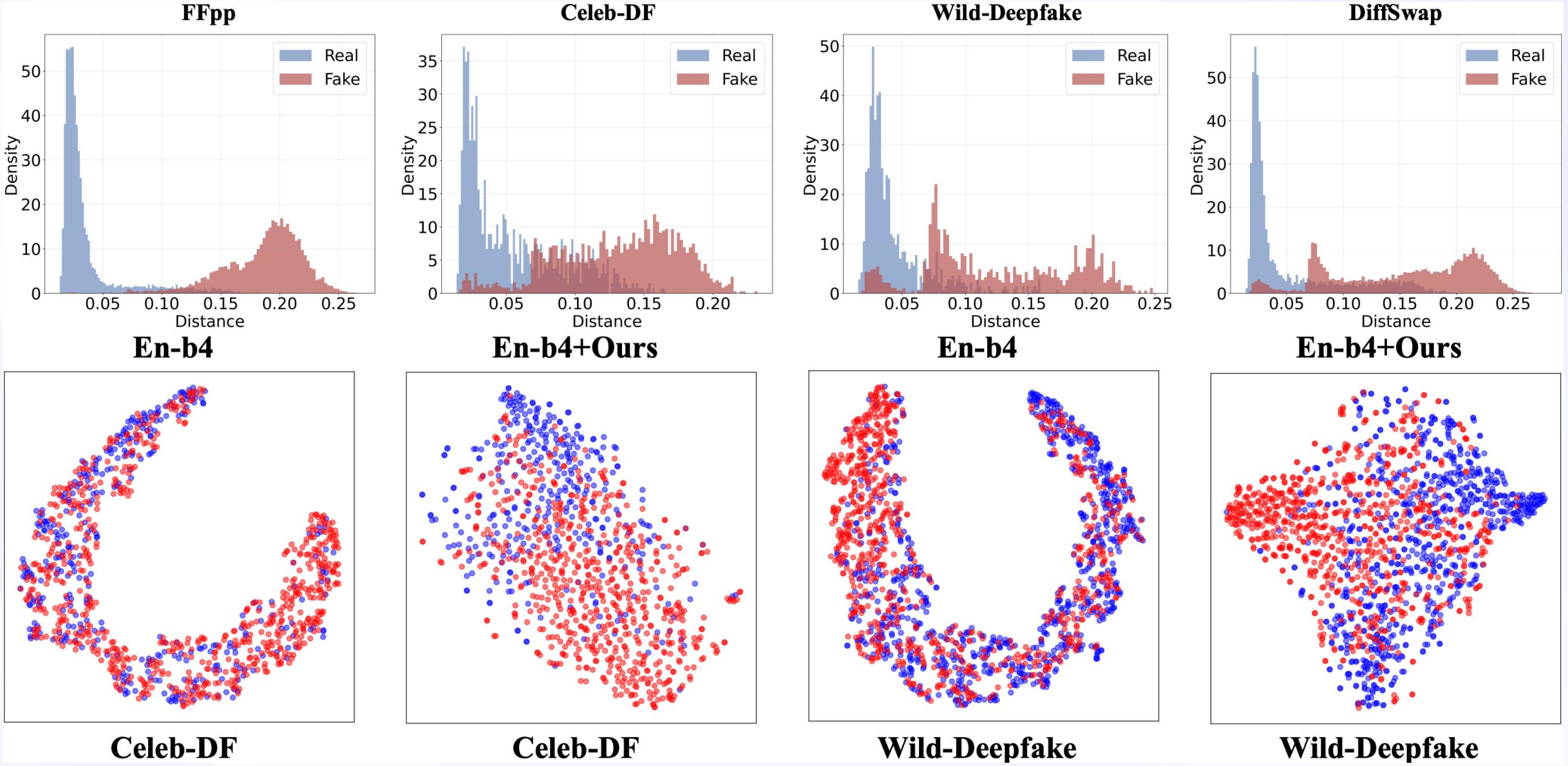
1）对于伪造图像而言，源图像与目标图像之间的差异显著。

2）对于真实图像而言，源图像与目标图像之间的差异甚微。



Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- 成效



Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- 成效

成效的探讨都是在训练过程中的：

1）特征过滤模块与权重模块通过剥离伪造图像中源图像与目标图像的特征以及为源图像和目标图像的特征作动态权重划分，提升了Encoder对伪造图像中源图像和目标图像的解耦能力；

2）Guide-moudle 通过提取和总结上述特征并与 Stable-diffusion 共同生成源图像和目标图像，使得Encoder更能学习到伪造图像中源图像与目标图像的特征，进而更理解伪造图像的语义特征。

3）DiffusionFake 训练框架的使用提升了 Encoder 对伪造图像其中的源图像与目标图像之间的不可见特征的分析能力和分类头的回归效果。

通过借助了 DiffusionFake 框架训练了 Encoder 和 分类头，进而使得模型获得了大量先验知识，使得其在最后推理过程中代价变低，有一些蒸馏模型的意思。

Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

DiffusionFake -- 可以改动的地方

1. 换一个更牛的扩散模型?

2. Encoder、Feature-Moudle、Weight-Moudle、Guide-Module 的结构?

3. 用这个插件训练一个其他的模型? （可尝试的空间最大）

Chen S, Yao T, Liu H, et al. DiffusionFake: Enhancing Generalization in Deepfake Detection via Guided Stable Diffusion[J]. Advances in Neural Information Processing Systems, 2025, 37: 101474-101497.

后续计划

尝试仿真 DiffusionFake 源代码

下周计划

01

基于 DiffusionFake 提出自己的模型

后续计划

02

谢　谢！