1 REVISED FIG.1

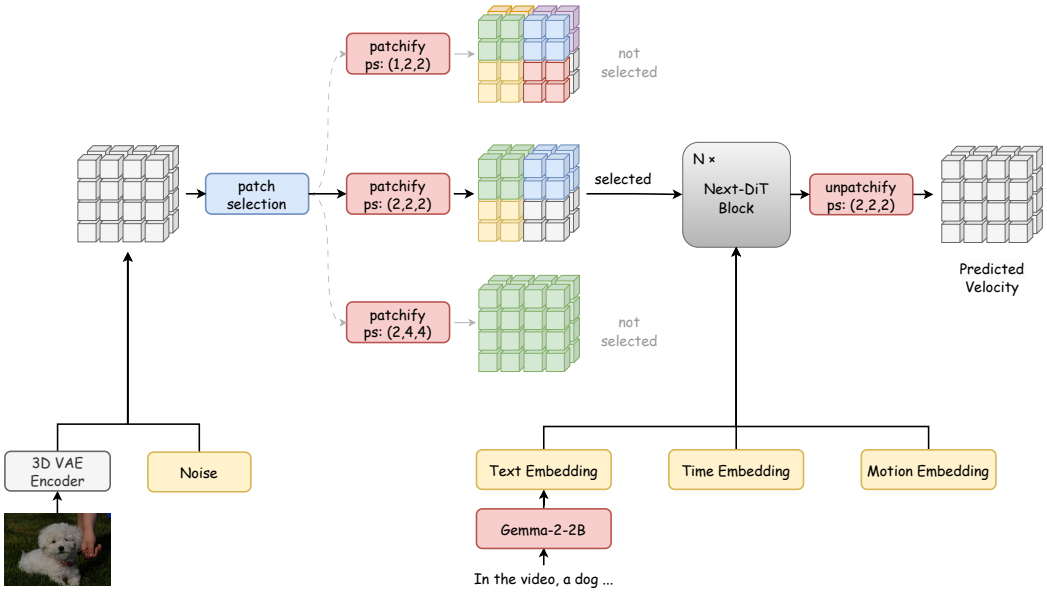

Figure 1: The revised version focuses more on multi-patchification and motion conditioning, and weakens the structure illustration of Next-DiT block, which is not the contribution of this paper.

2 ADDITIONAL TABLES

Table 1: Ablation study on time shift. Single/Multi represents single/multi-scale Inference. VBench Total Score is reported.

Time Shift	Res512&FPS16 (Single)	Res512&FPS16 (Multi)	Res960&FPS24 (Multi)
4	82.75	82.69	82.44
8	82.99	82.94	82.67
16	82.92	82.94	83.08
24	82.74	82.86	83.12
32	82.62	82.71	82.93

Table 2: Summarization of time shift values we use for training.

Stage	Type	Patch=1x2	Patch=2x2	Patch=2x4
1	Image (Res256)	1	-	-
2	Image (Res256)	1	-	-
	Video (Res256&FPS8)	2	4	6
3	Image (Res512)	2	-	-
	Video (Res512&FPS16)	4	8	12
4	Image (Res960)	4	-	-
	Video (Res960&FPS16)	8	16	24

Table 3: Ablation study on denoising steps. Setting: Res=512&FPS=16.

Steps	Time Shift=8		Time Shift=16	
	Single Scale	Multi Scale	Single Scale	Multi Scale
30	82.30	82.24	82.66	82.57
50	82.82	82.79	82.96	82.95
70	82.99	82.94	82.92	82.94