

split attention probability for all attention heads for pre-norm

